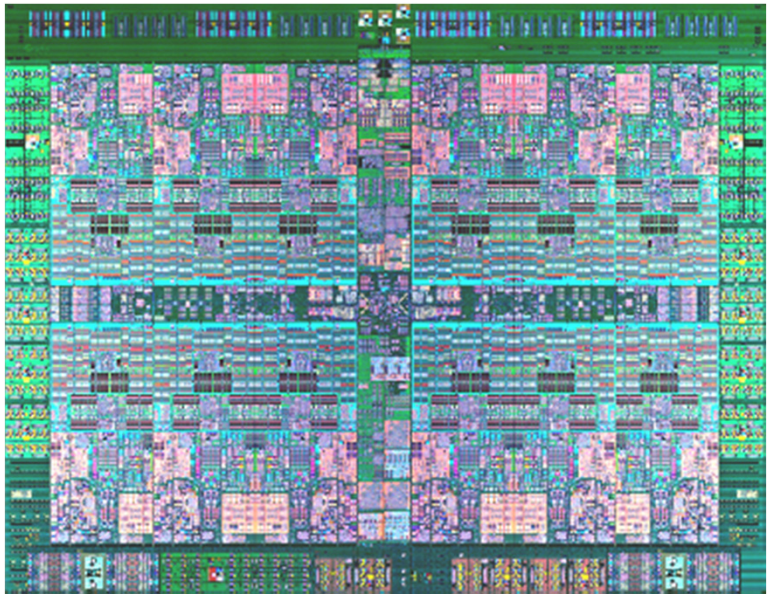
Welcome To

ISSCC 2014
Session 5

Processors

POWER8™: A 12-Core Server-Class Processor in 22nm SOI with 7.6Tb/s Off-Chip Bandwidth



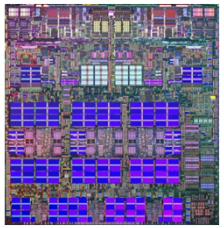
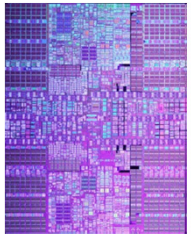
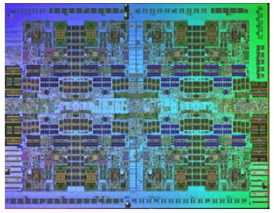
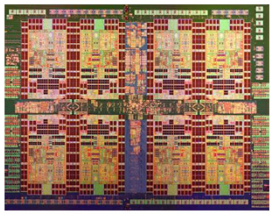
Eric Fluhr

Joshua Friedrich, Daniel Dreps, Victor Zyuban, Gregory Still, Christopher Gonzalez, Allen Hall, David Hogenmiller, Frank Malgioglio, Ryan Nett, Jose Paredes, Juergen Pille, Donald Plass, Ruchir Puri, Phillip Restle, David Shan, Kevin Stawiasz, Zeynep Toprak Deniz, Dieter Wendel, Matt Ziegler, Howard Smith

Outline

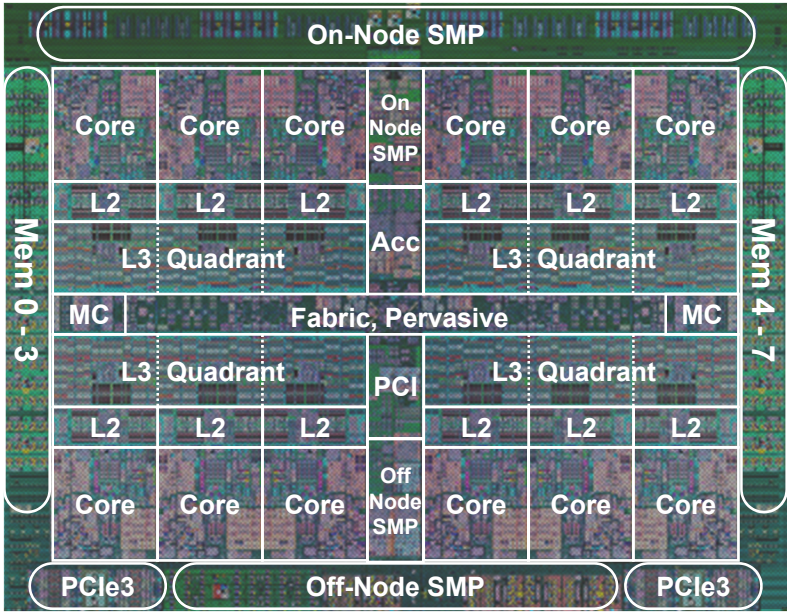
- Data optimized design
  - Technology
  - Highly threaded, wide execution core
  - High bandwidth nest
- Circuit optimizations
  - Arrays, Clocking, and IOs
- Power Management & Reduction
- Design methodology
- Lab data

“Recent” POWER History

	POWER5 *2004*	POWER6 *2007*	POWER7 *2010*	POWER7+ *2012*	
					
Technology	130nm SOI	65nm SOI	45nm SOI eDRAM	32nm SOI eDRAM	22nm SOI eDRAM
Compute					
Cores	2	2	8	8	12
Threads	SMT2	SMT2	SMT4	SMT4	SMT8
Caching					
On-chip	1.9MB	8MB	2 + 32MB	2 + 80MB	6 + 96MB
Off-chip	36MB	32MB	None	None	128MB
Bandwidth					
Sust. Mem.	15GB/s	30GB/s	100GB/s	100GB/s	230GB/s
Peak I/O	6GB/s	20GB/s	40GB/s	40GB/s	64GB/s

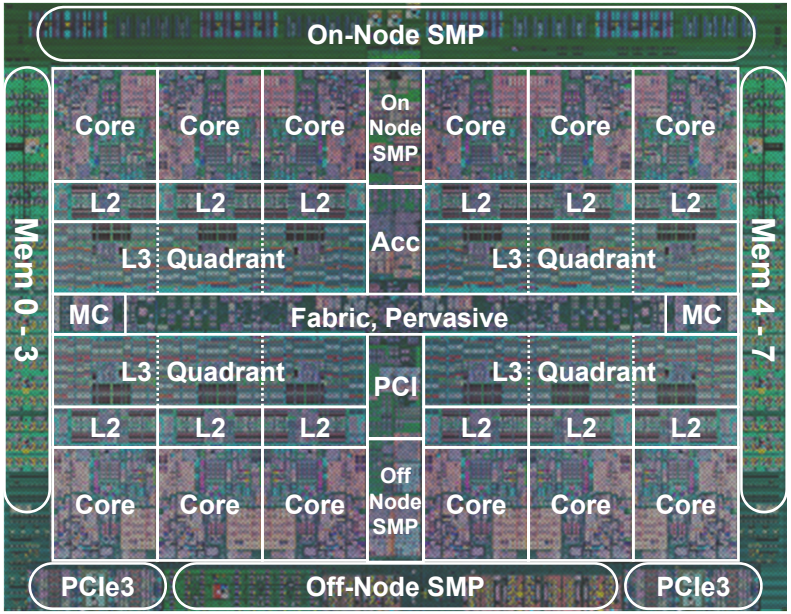
POWER8 Chip Overview

- Up to 2.5x socket perf vs. POWER7+
- 649mm² die size, 4.2B transistors

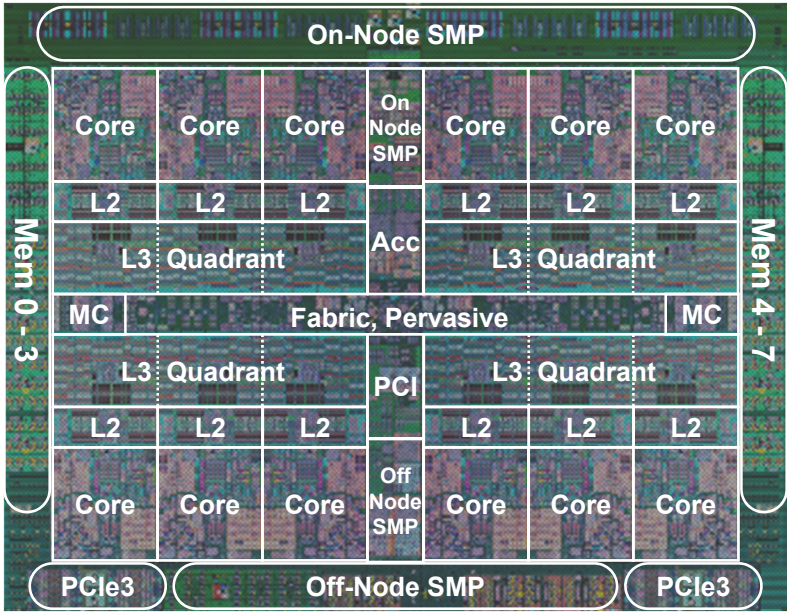


POWER8 Chip Overview

- Up to 2.5x socket perf vs. POWER7+
- 649mm² die size, 4.2B transistors
- 12 high-performance cores
- Large Caches
  - L2: 512KB private SRAM per core
  - L3: 96MB shared eDRAM w/ 8MB “fast access” partition per core
  - L4: Up to 128MB, located on memory buffer chips



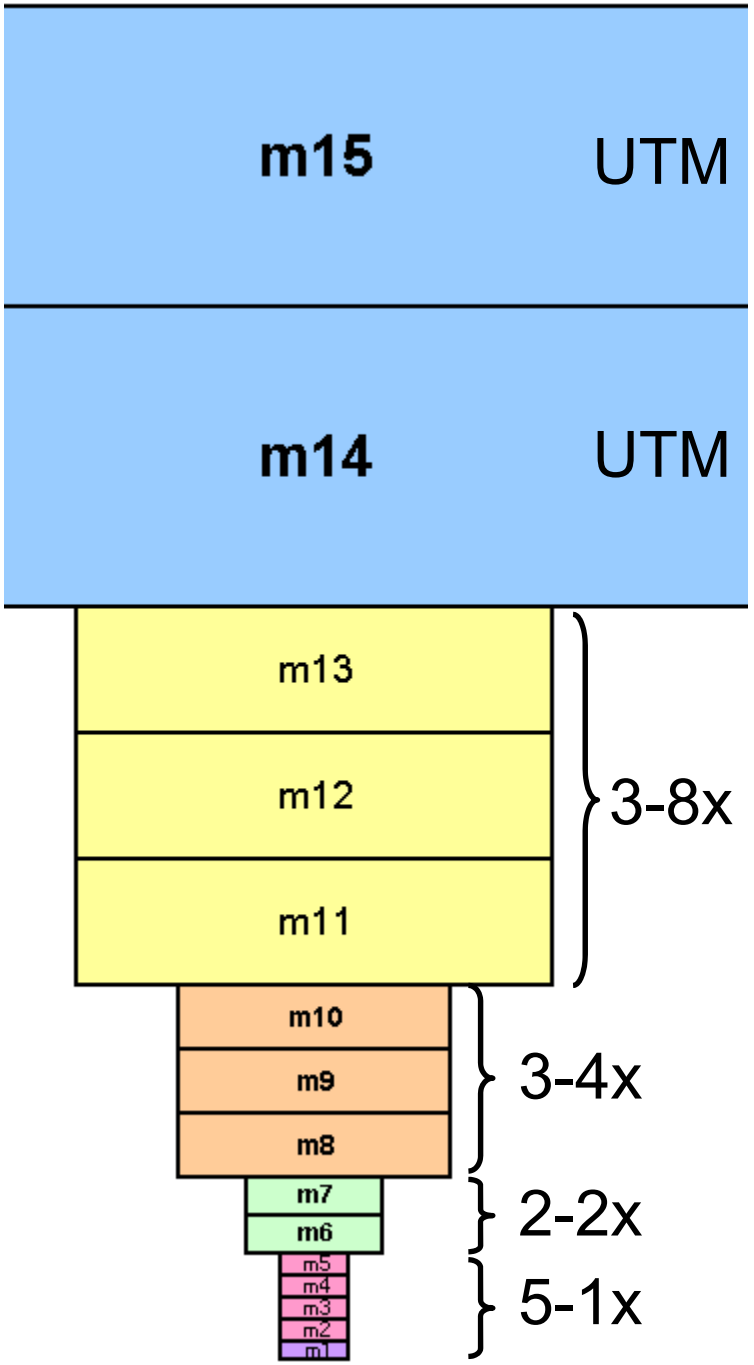
POWER8 Chip Overview



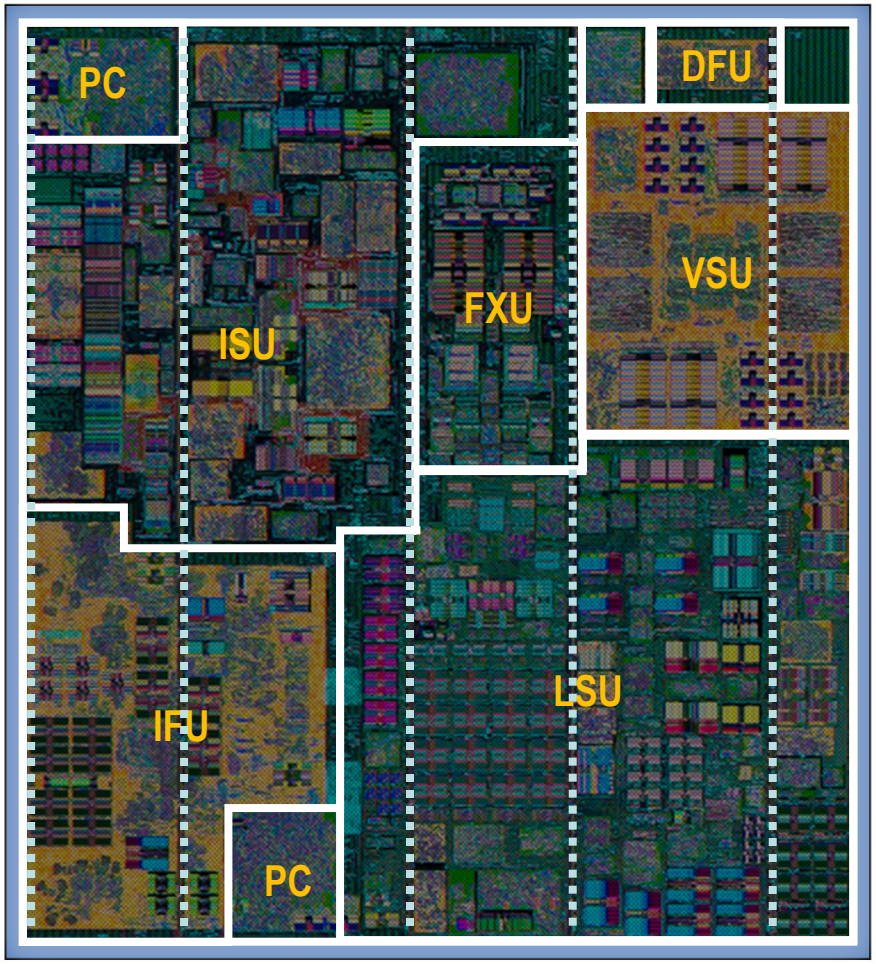
- Up to 2.5x socket perf vs. POWER7+
- 649mm² die size, 4.2B transistors
- 12 high-performance cores
- Large Caches
  - L2: 512KB private SRAM per core
  - L3: 96MB shared eDRAM w/ 8MB “fast access” partition per core
  - L4: Up to 128MB, located on memory buffer chips
- 4 High Throughput I/O interfaces
  - Memory, On-Node SMP, Off-Node SMP, PCIe Gen3
- CAPI: open infrastructure for off-chip, memory-coherent accelerators

POWER8 Technology

- 22nm SOI
- 15 layer BEOL:
  5-1x, 2-2x, 3-4x, 3-8x, 2-UTM
- 3-Vt thin-oxide logic transistors for power optimization
- Multiple thick-oxide transistors (for I/O and analog support)
- 3 app-optimized SRAM cells:
  - 0.160um2 6T perf-oriented
  - 0.144um2 6T perf-density balance for directories/L2
  - 0.192um2 8T multi-port
- Technology eDRAM cell: 0.026um2



POWER8 Core



Enhanced Micro Architecture

- Increased Execution Bandwidth, +4 units
- SMT 8
- 64KB L1 D-Cache, 32KB 8-way I-Cache
- 64B Cache Reload
- 4KB TLB
- Transactional Memory

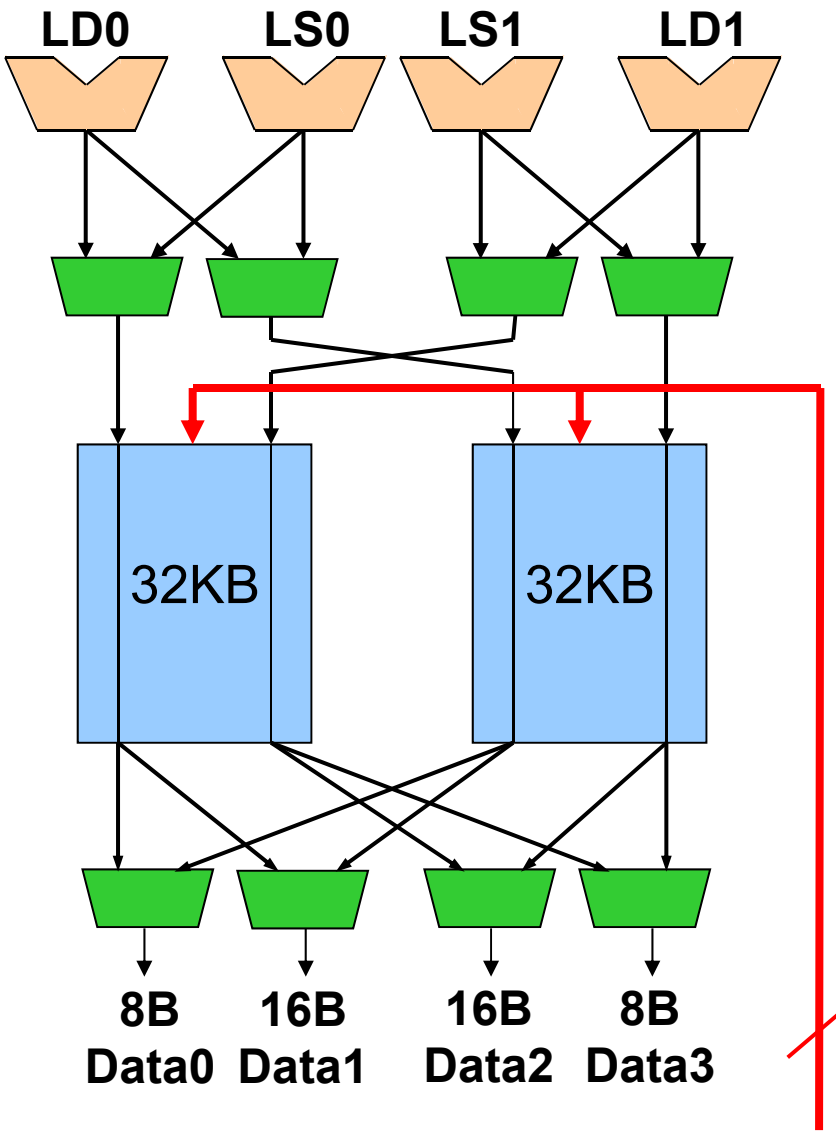
Arrays/Register Files

- 2 CAM & 6 SRAM Topologies
- 31 Multi-ported Register Files for Queuing & Architected Registers

Power Management

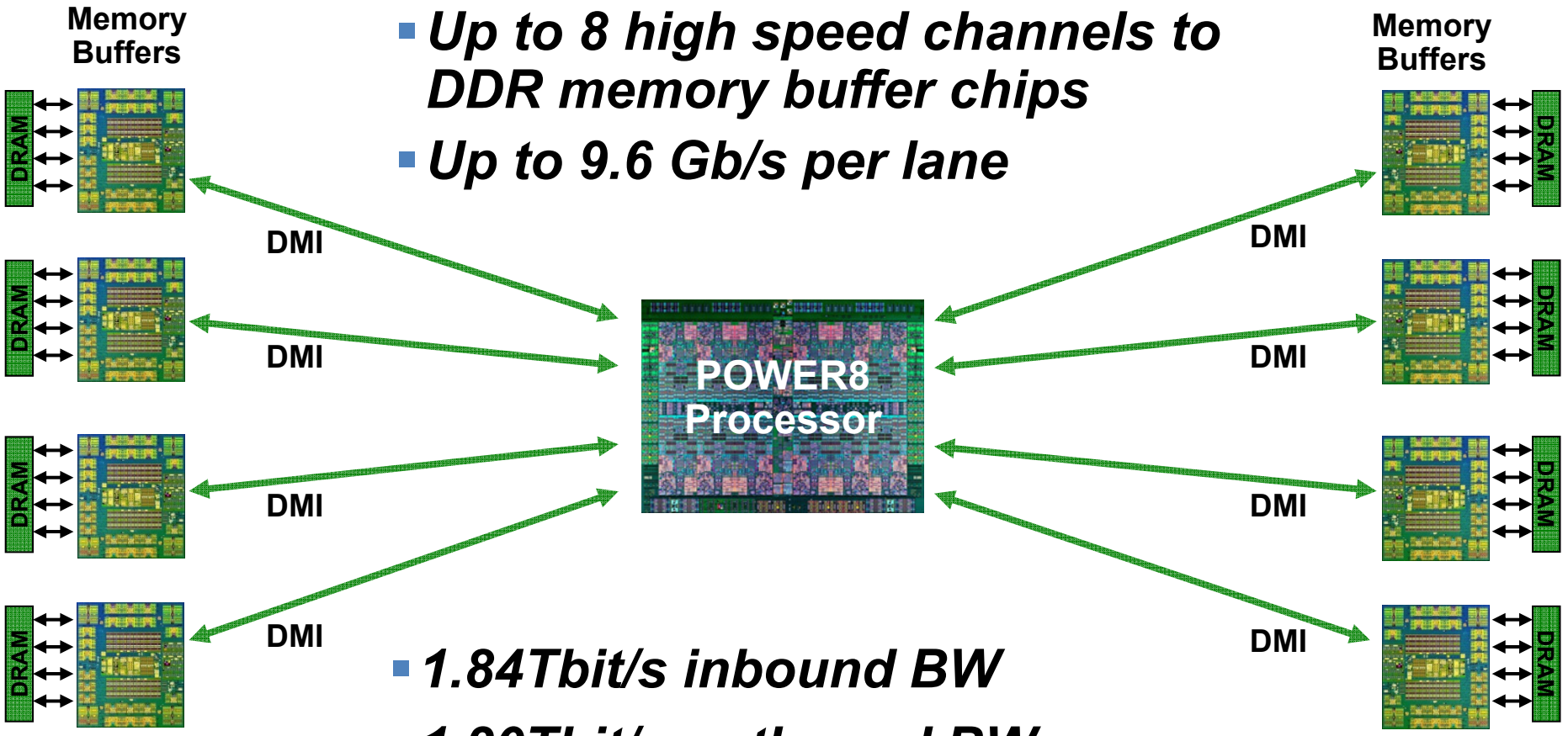
- Power Gating & Voltage Regulation in 5 columns
- 1 Thermal Diode
- 3 Digital Thermal Sensors
- 3 Critical Path Monitors

High Bandwidth L1 Data Cache



- 64KB banked design
- 2 Load x 8B
- 2 Load/Store x 16B
- 40B per cycle peak load
- 16B per cycle peak store
- 64B reload / cycle from L2

POWER8 Memory Bandwidth

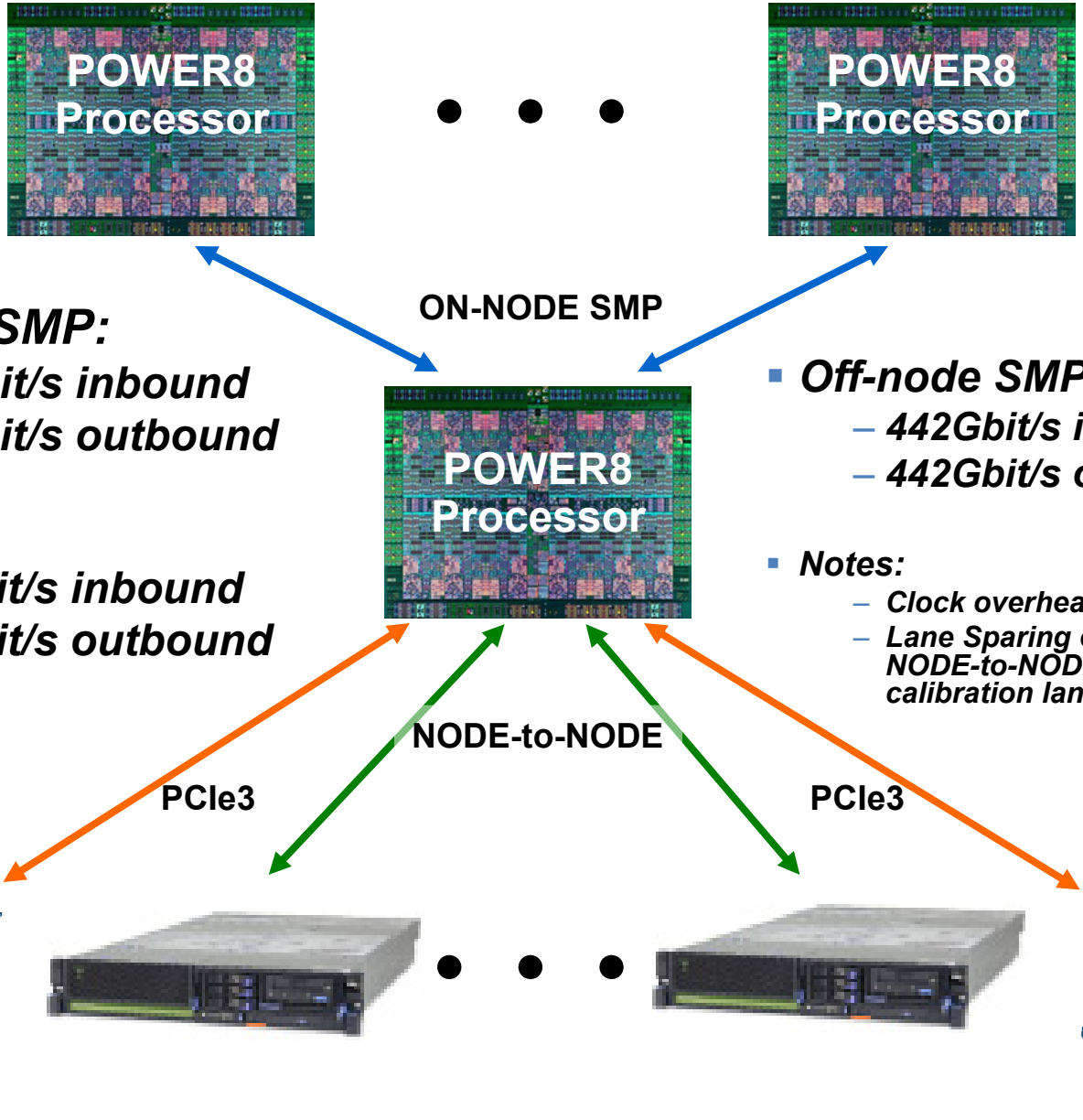


- ***Up to 8 high speed channels to DDR memory buffer chips***
- ***Up to 9.6 Gb/s per lane***

- ***1.84Tbit/s inbound BW***
- ***1.30Tbit/s outbound BW***
- ***Total chip capacity up to 1TB***

- ***Notes: clock overhead removed; includes lane sparing and an additional calibration lane***

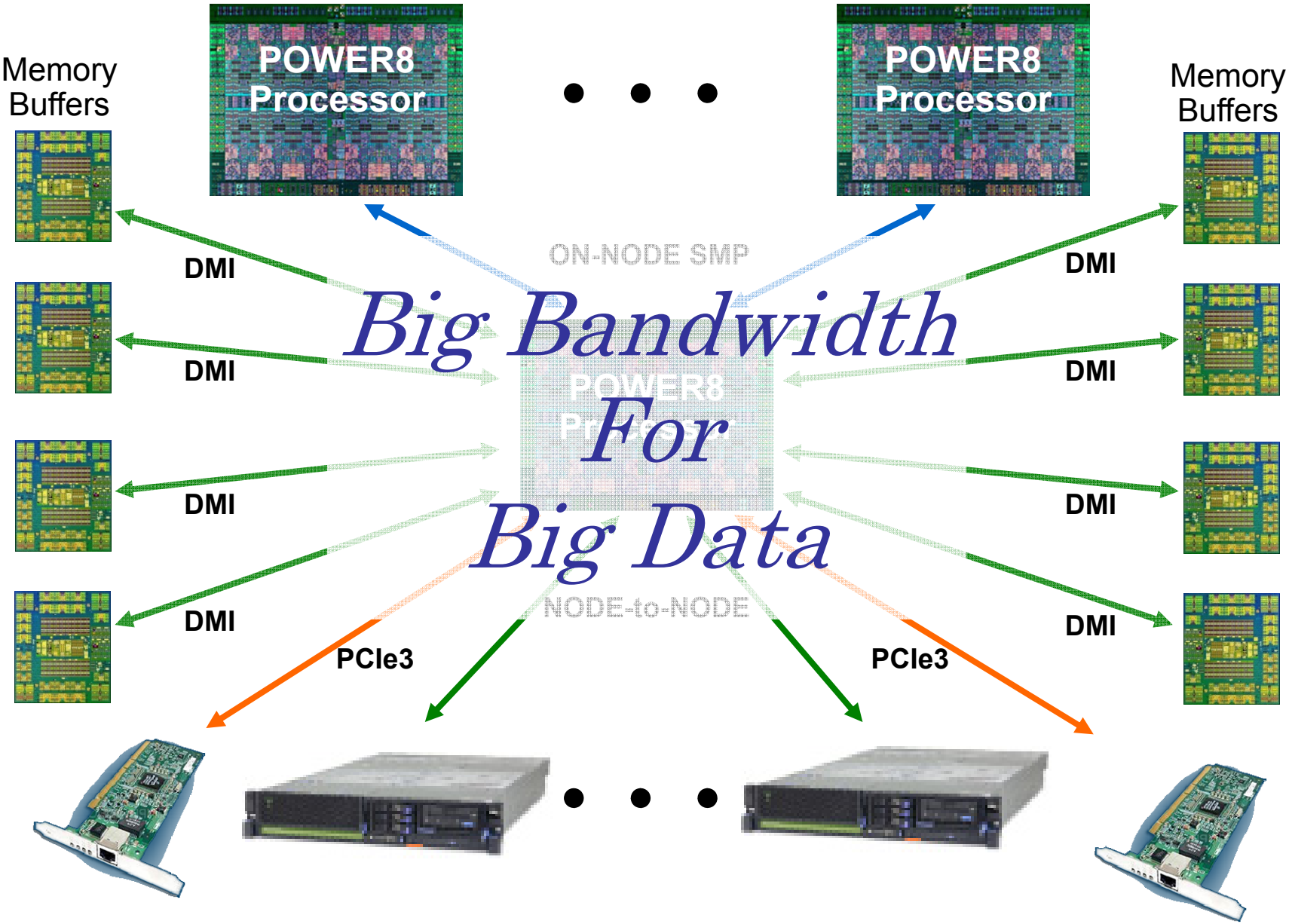
POWER8 SMP & PCIe Bandwidth



- ***On-node SMP:***
  - ***1.54Tbit/s inbound***
  - ***1.54Tbit/s outbound***
- ***PCI BW:***
  - ***256Gbit/s inbound***
  - ***256Gbit/s outbound***

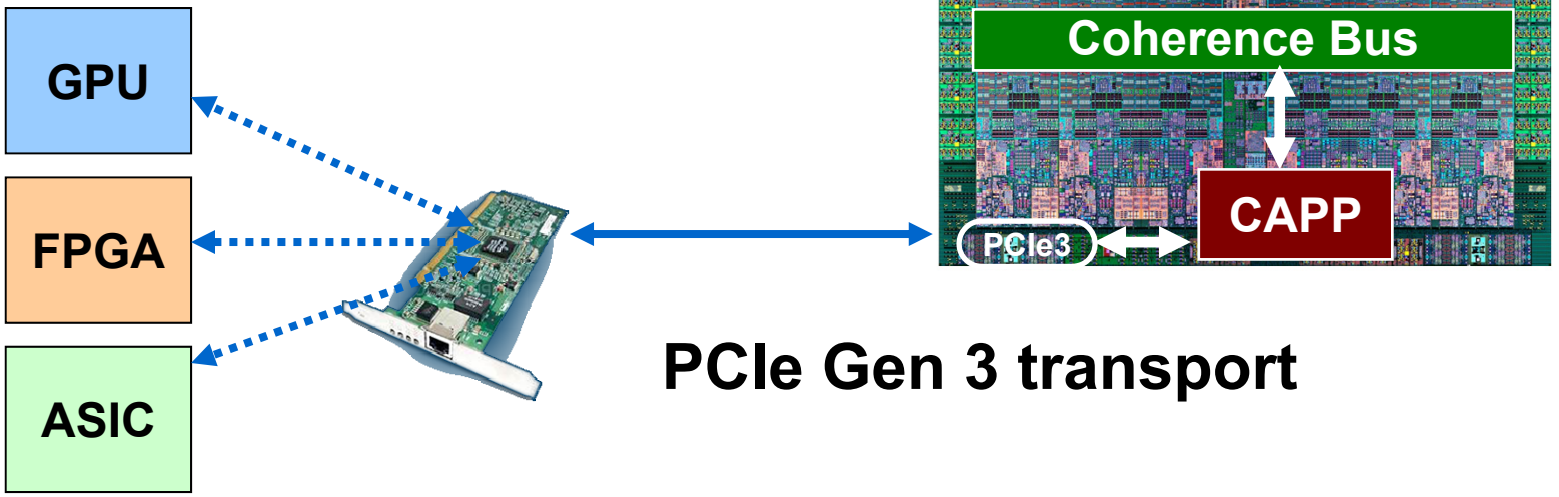
- ***Off-node SMP:***
  - ***442Gbit/s inbound***
  - ***442Gbit/s outbound***
- ***Notes:***
  - ***Clock overhead removed***
  - ***Lane Sparing exists in all but PCI; NODE-to-NODE includes calibration lane in the accounting***

Combined I/O Bandwidth = 7.6Tb/s



POWER8 CAPI

- **Coherent Accelerator Processor Interface**
- **Hardware Managed Cache Coherence:**
  Lowers Latency over IO communication model
- **Virtual Addressing:**
  Removes OS & device driver overhead
- **Customizable Hardware Application Accelerators**

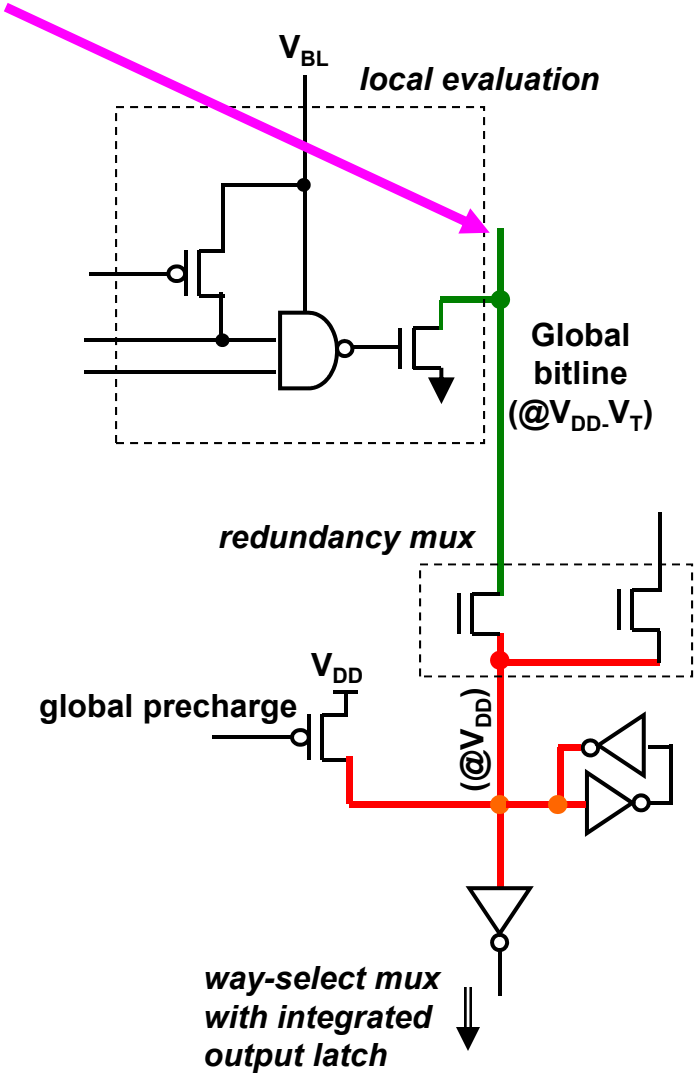


Outline

- Data optimized design
  - Technology
  - Highly threaded, wide execution core
  - High bandwidth nest
- Circuit optimizations
  - Arrays and Clocking
- Power Management & Reduction
- Design methodology
- Lab data

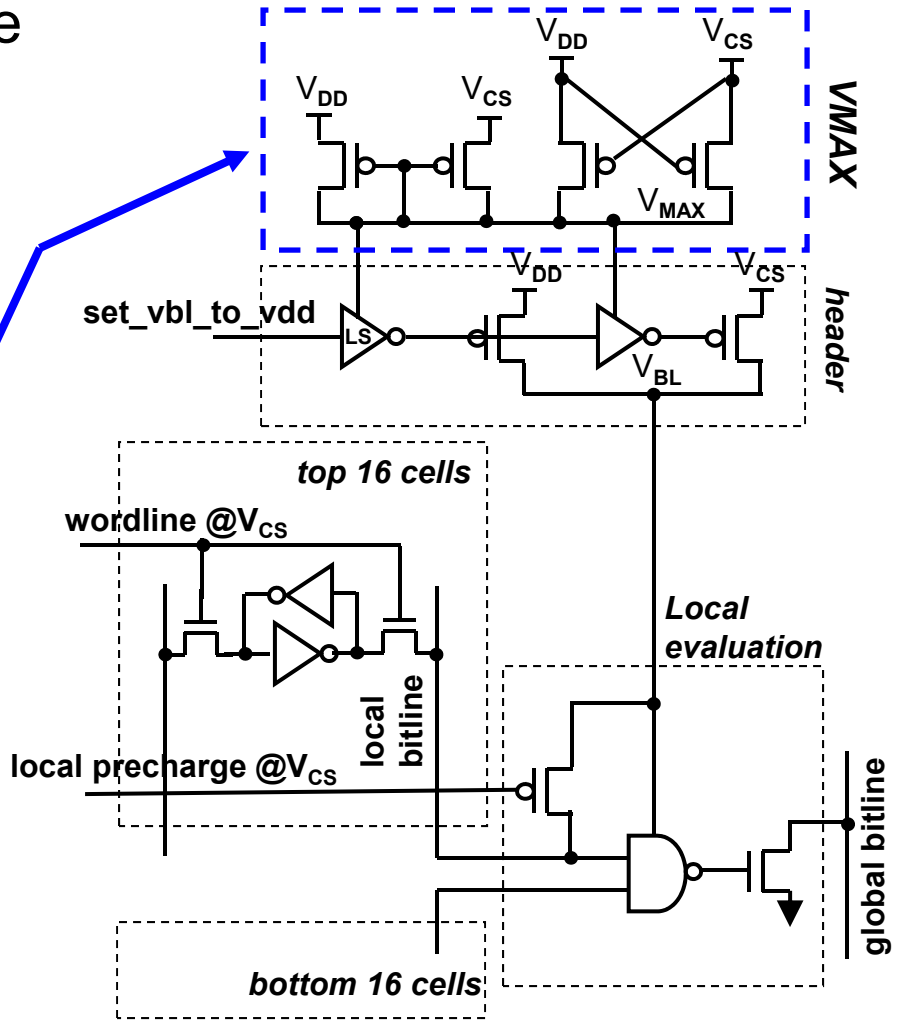
SRAM Power Savings

- Global bitline restored to reduced voltage V_{DD}-V_T
  - 20% AC power savings
- Smart way select prediction to reduce restore power
- Early and late wordline gating features
- Wordline driver header devices
  - 16% DC power savings
- Output socket buffer concept: driver size tuned load of each instance

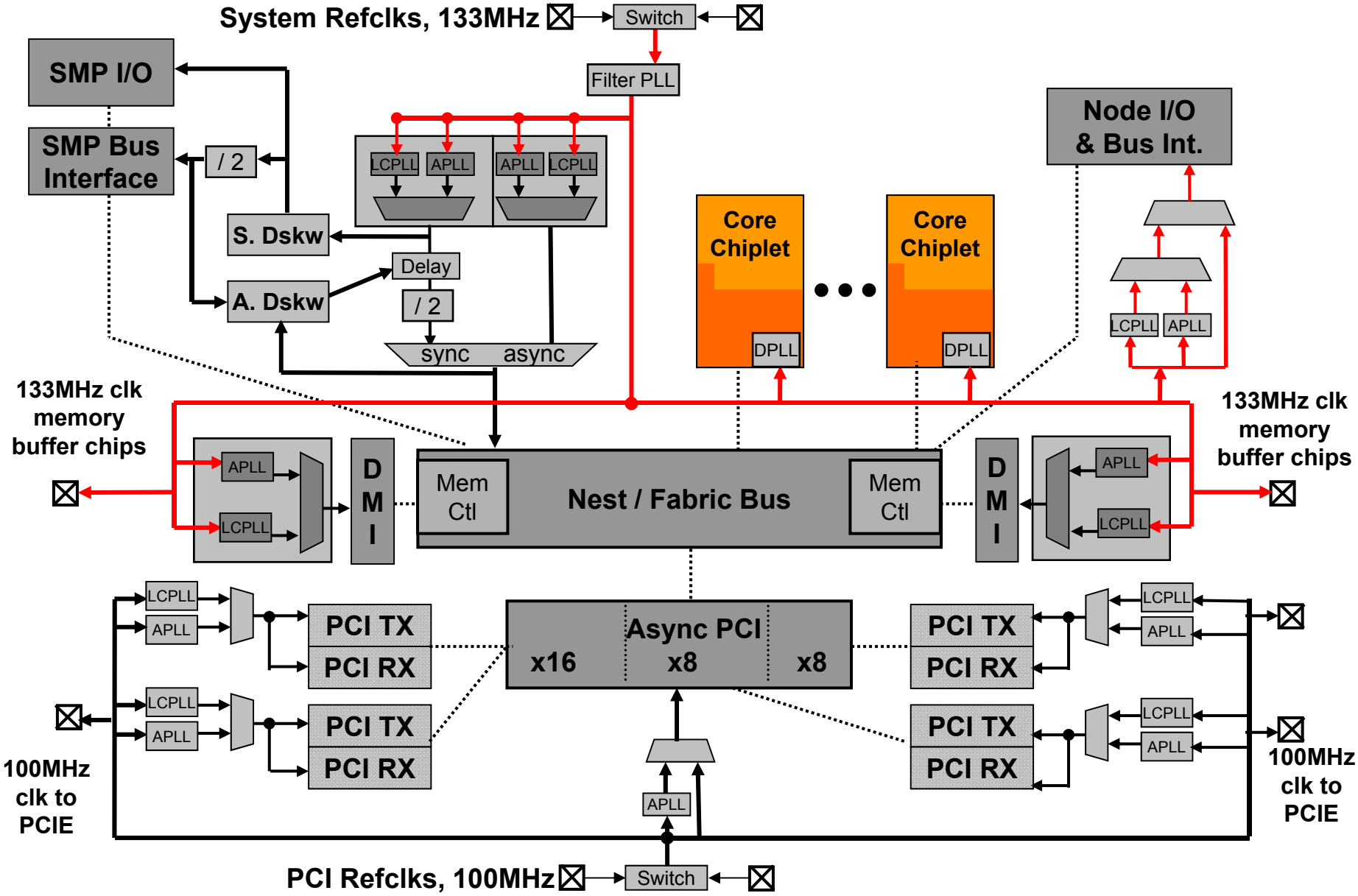


Performance SRAMs: Dual-mode BL Voltage

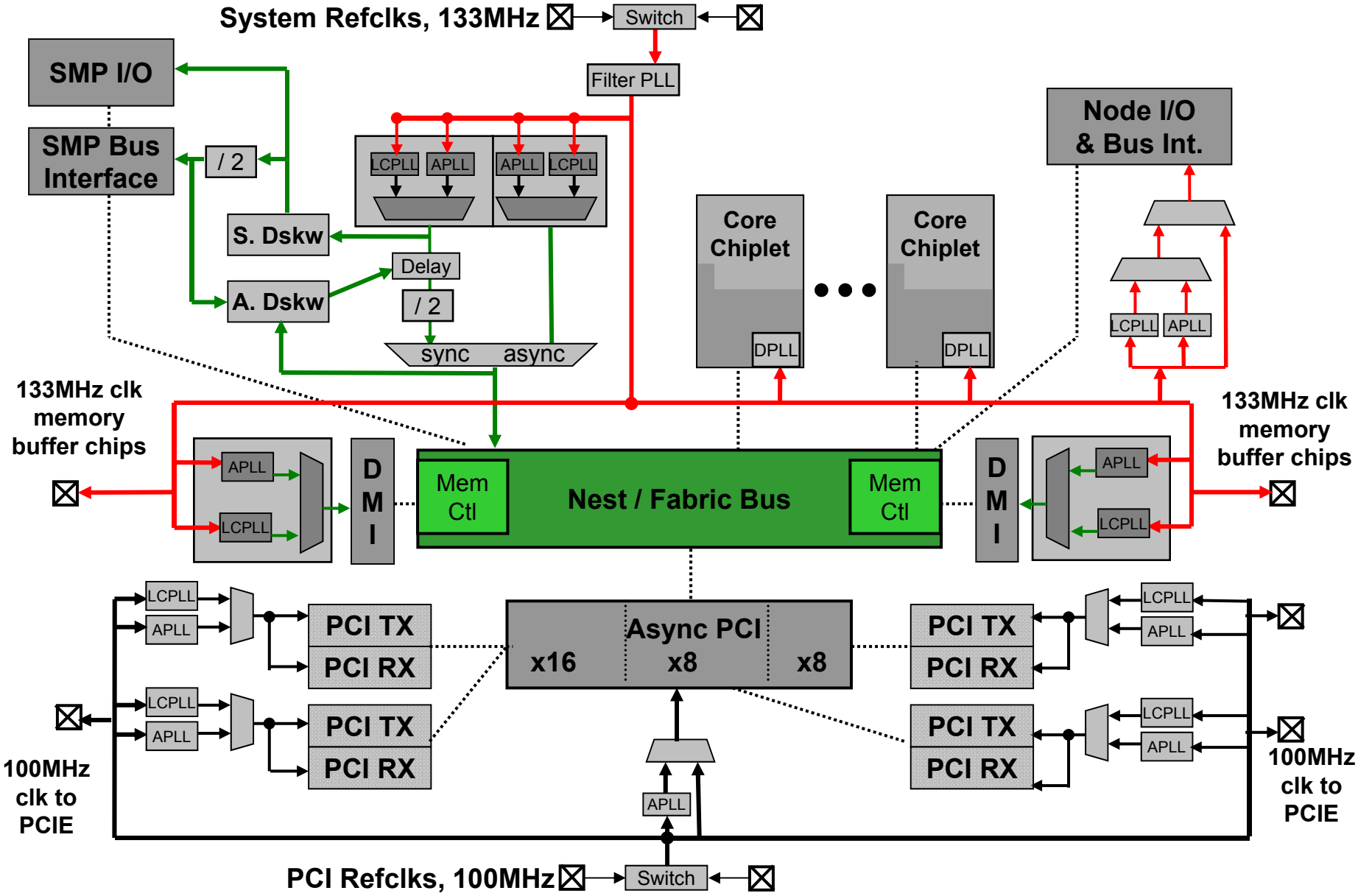
- Single-ended ripple-domino sense scheme
- Functional: $V_{CELL}=V_{BL}=V_{CS}$
  - Higher array voltage **V_{CS}** for stability
  - **V_{CELL}** & **V_{BL}** at same voltage to minimize voltage difference
- Test: $V_{CELL}=V_{CS}$, $V_{BL}=V_{DD}$
  - EOL margin test and characterization
- VMAX circuit to prevent any cross-circuit currents during power on sequence



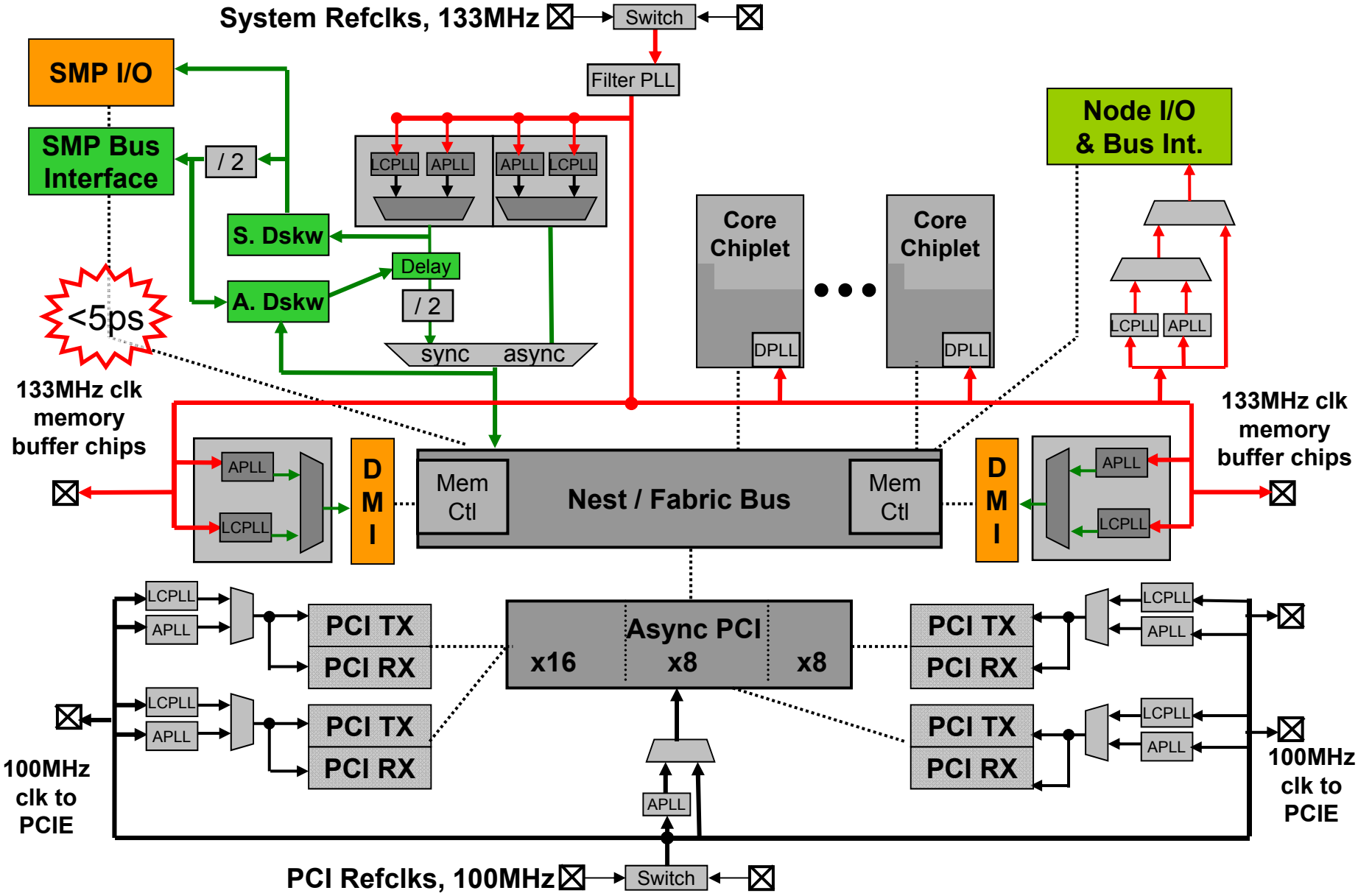
Clock Topology: 12 core/L2 with 12 L3 variable freq. meshes



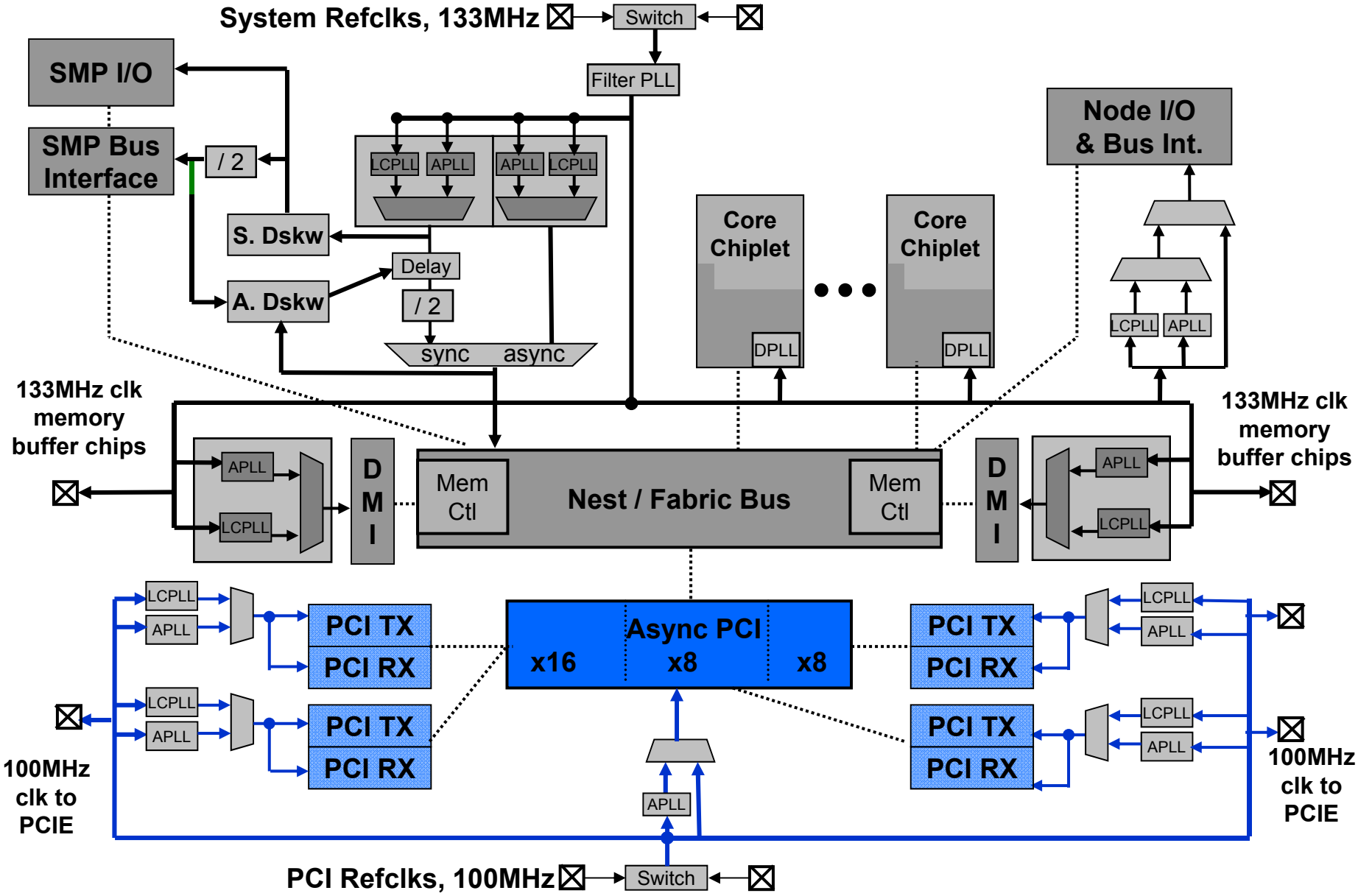
Clock Topology: nest/memory constant freq. mesh



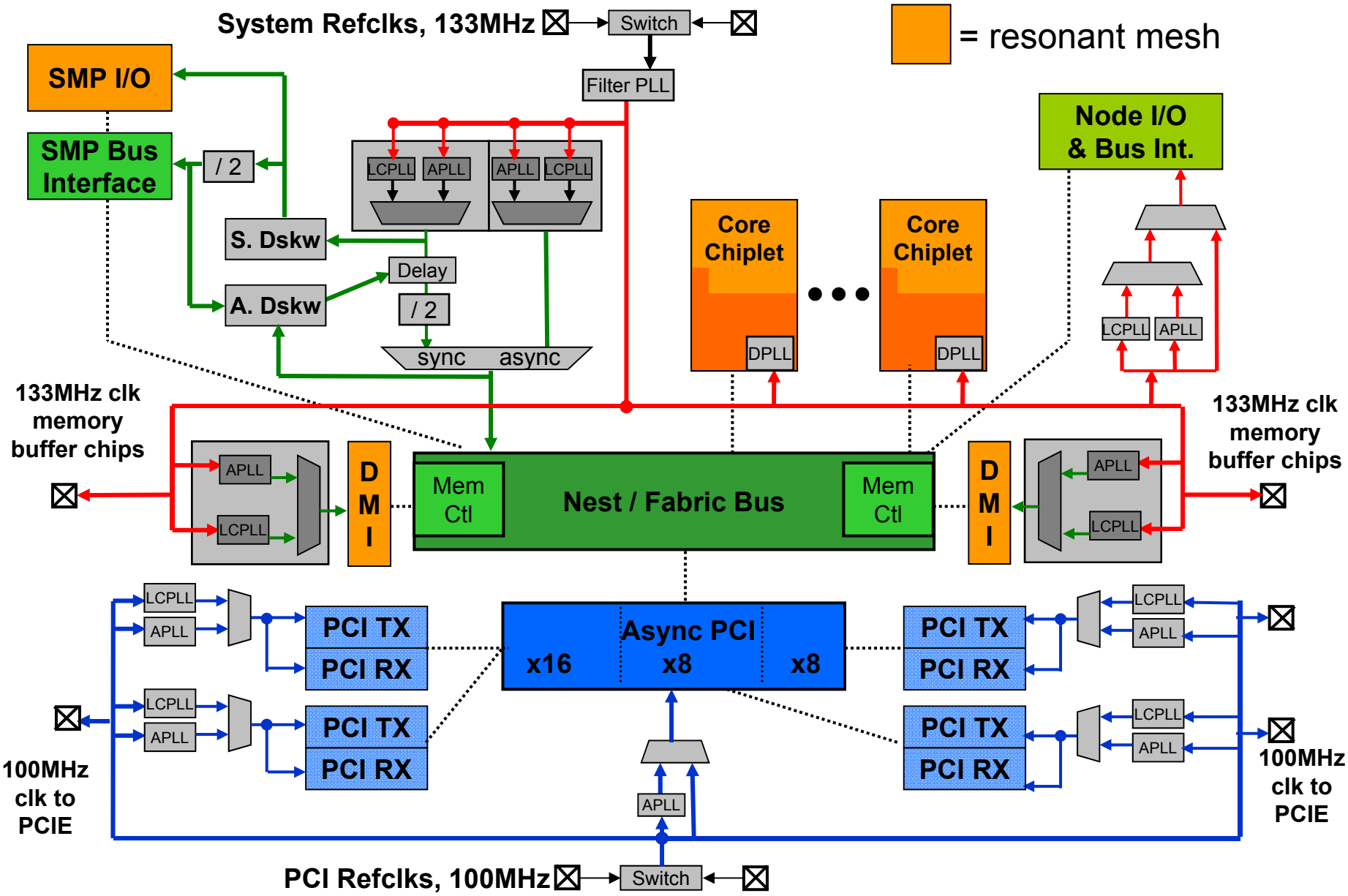
Clock Topology: on-node SMP, off-node SMP, & DMI



Clock Topology: PCI asynchronous



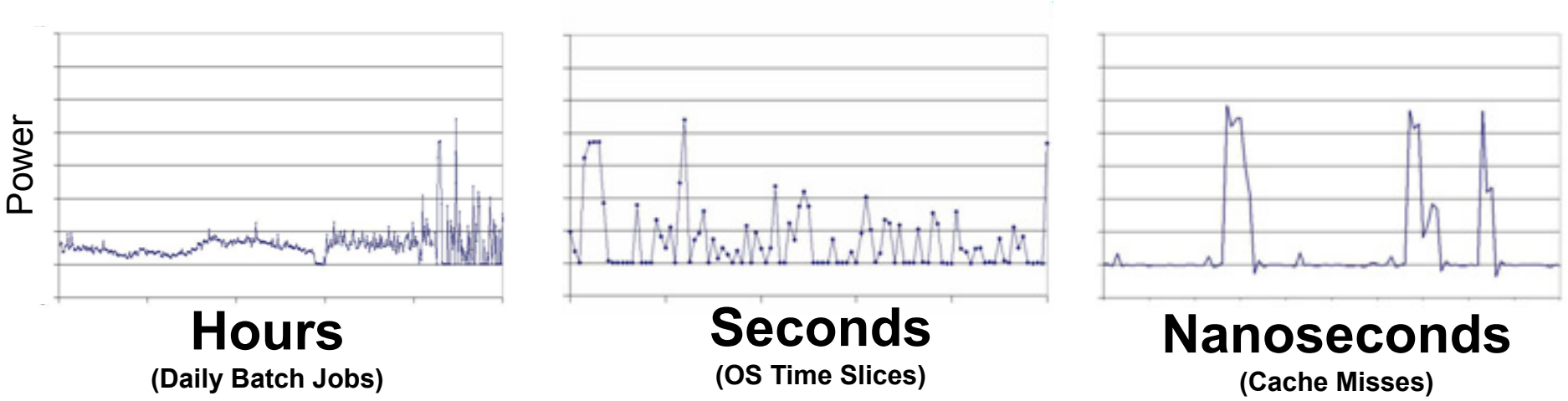
Clock Topology: 29 Domains



Outline

- Data optimized design
  - Technology
  - Highly threaded, wide execution core
  - High bandwidth nest
- Circuit optimizations
  - Arrays and Clocking
- Power Management & Reduction
- Design methodology
- Lab data

Exploiting processor inactivity



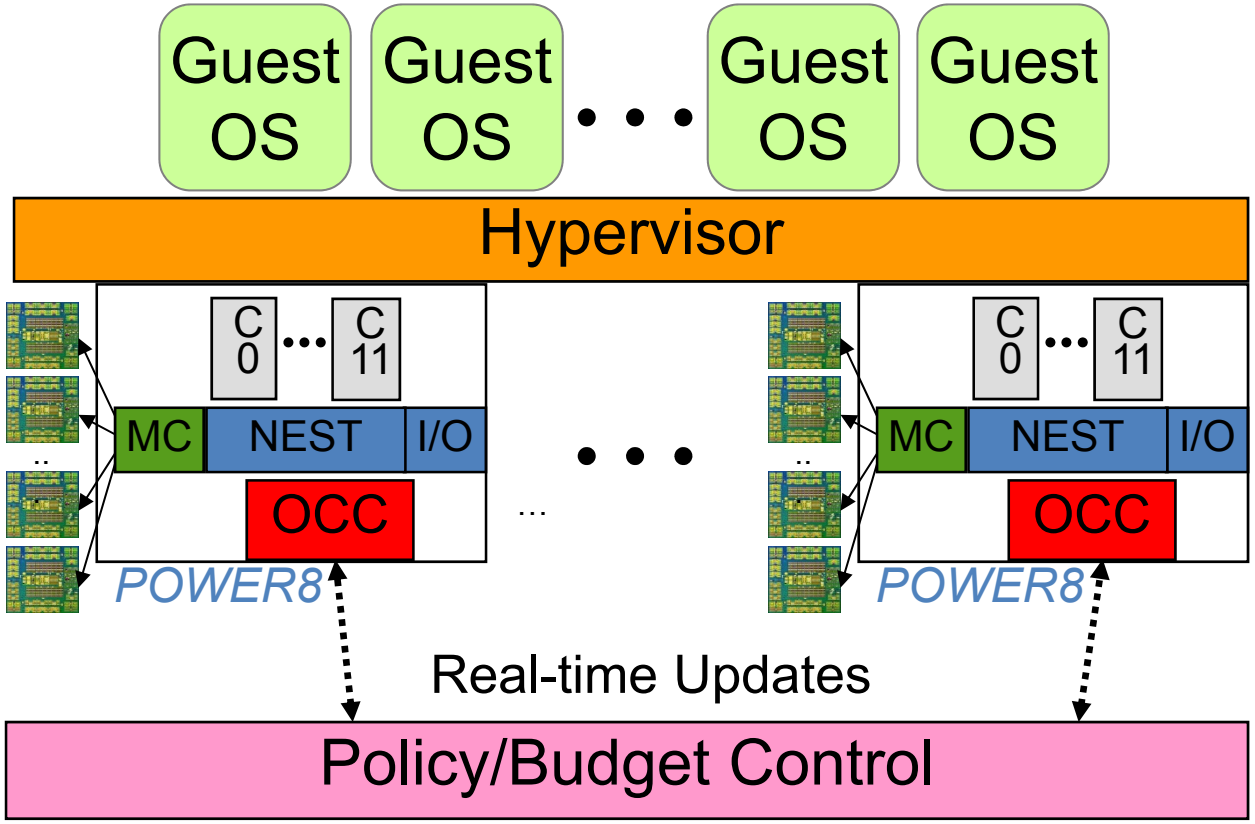
- **Power consumption varies at every time scale**
- **Key = sense and act in time**

POWER8 On-Chip Controller (OCC)

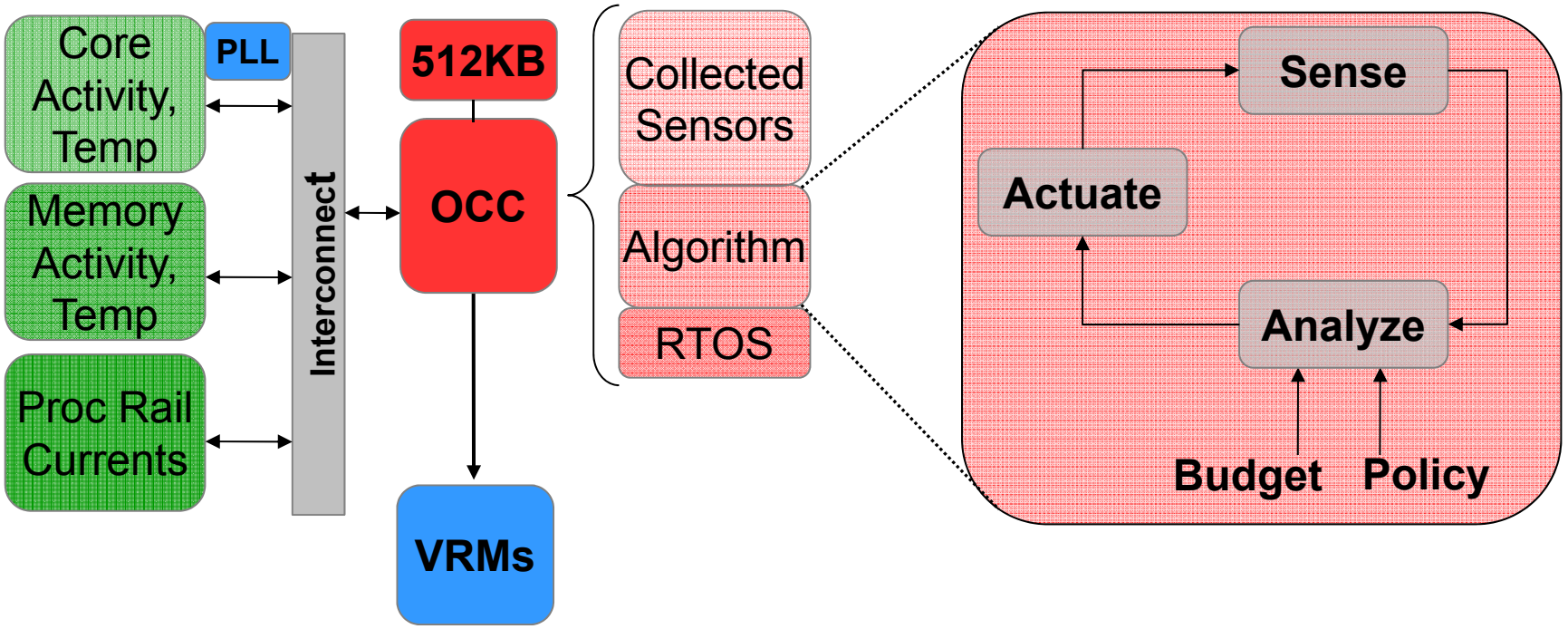
- Allows for fast, scalable monitoring and response
  - Independent of Hypervisor or Guest OS(s)

OR

  - In conjunction with Hypervisor interaction with Guest OS(s)

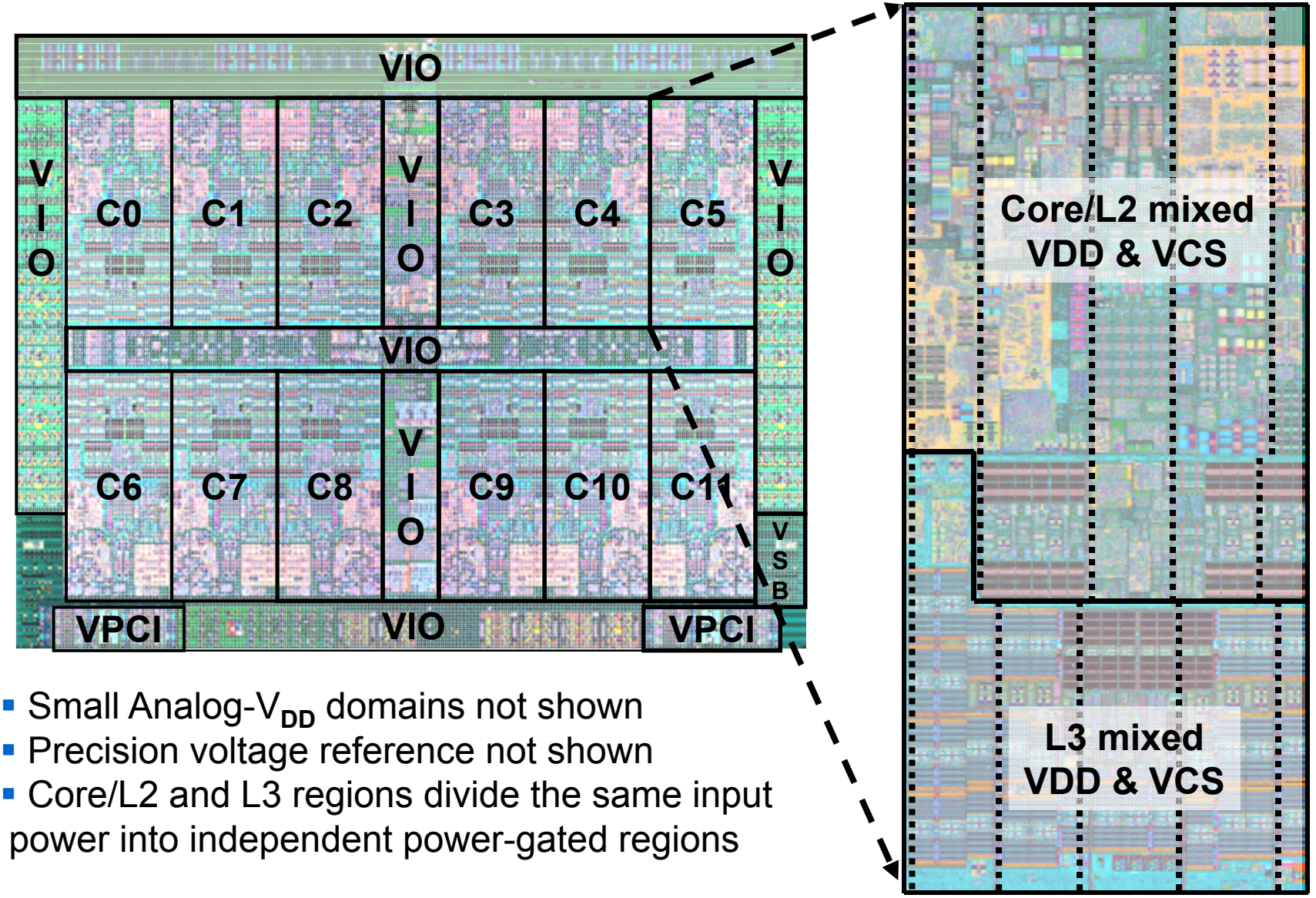


Faster Power Management == ideal for cloud!



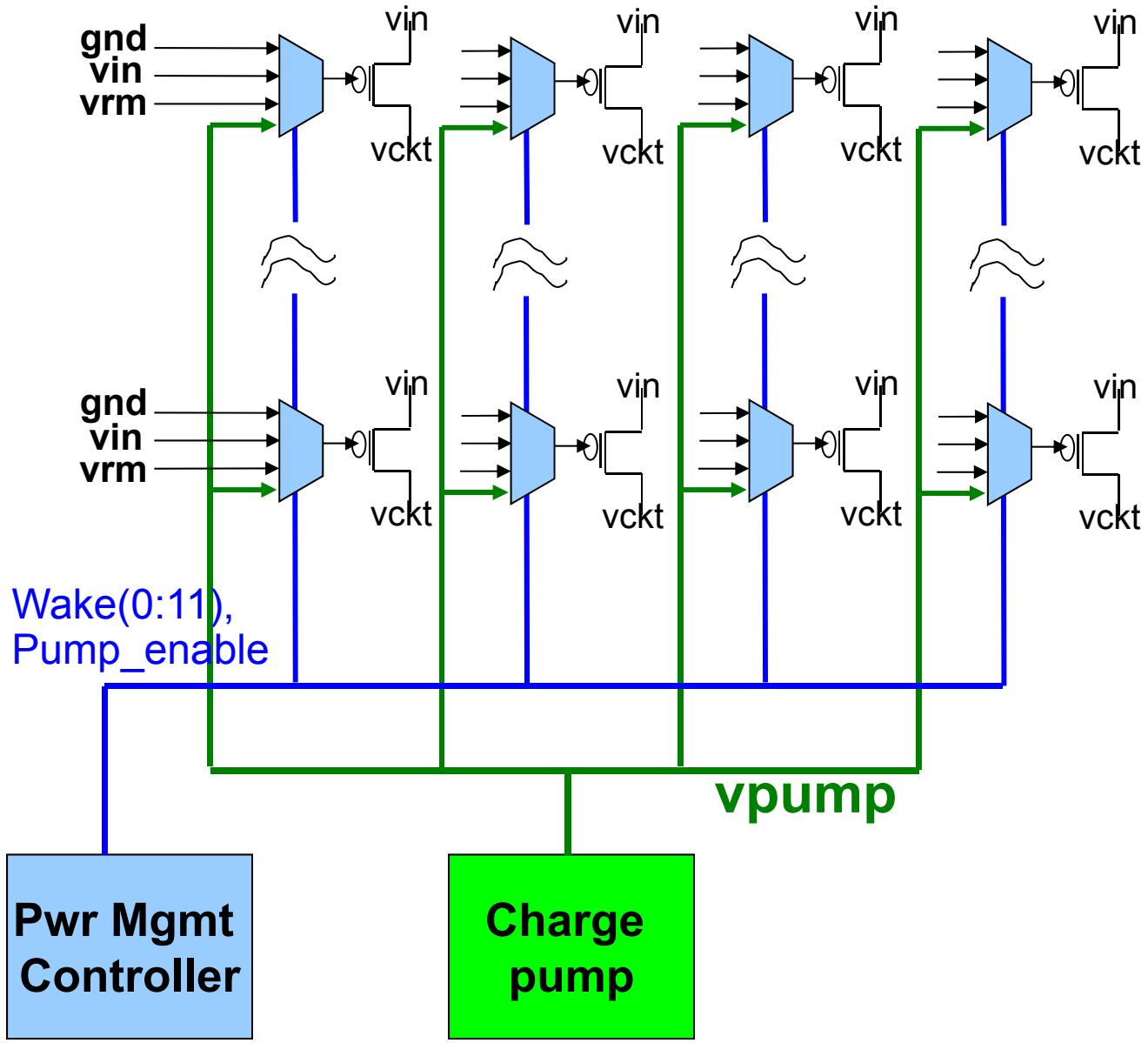
- OCC = full POWERPC 405 core with 512KB private memory
- Uses continuous running, real-time OS
- Monitors workload activity, chip temperature and current
- Adjusts frequency and voltage to optimize performance within system power and thermal constraints

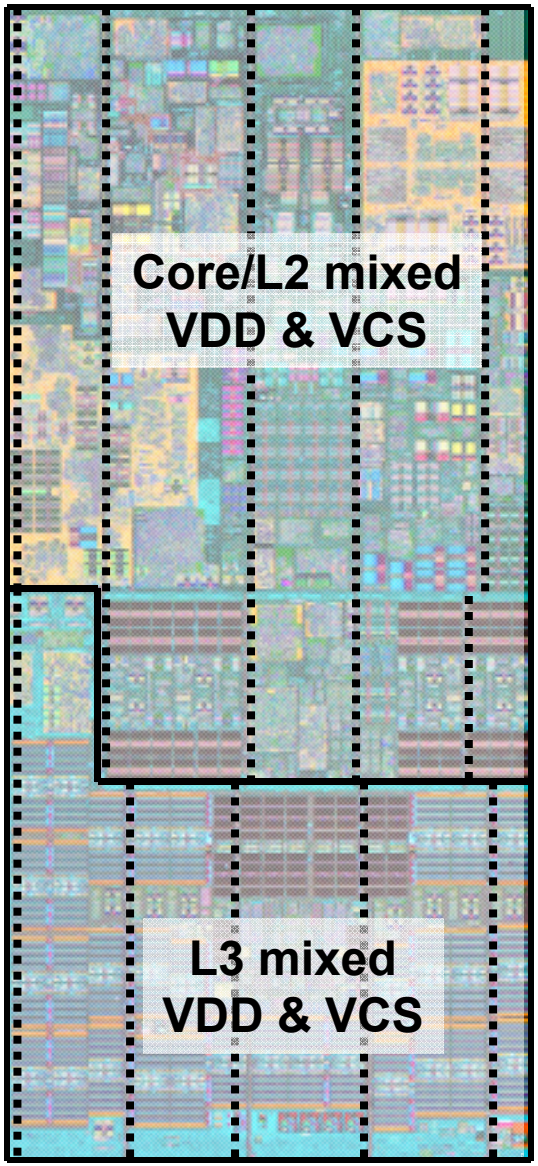
POWER8 Voltage Regions



- Small Analog-V_{DD} domains not shown
- Precision voltage reference not shown
- Core/L2 and L3 regions divide the same input power into independent power-gated regions

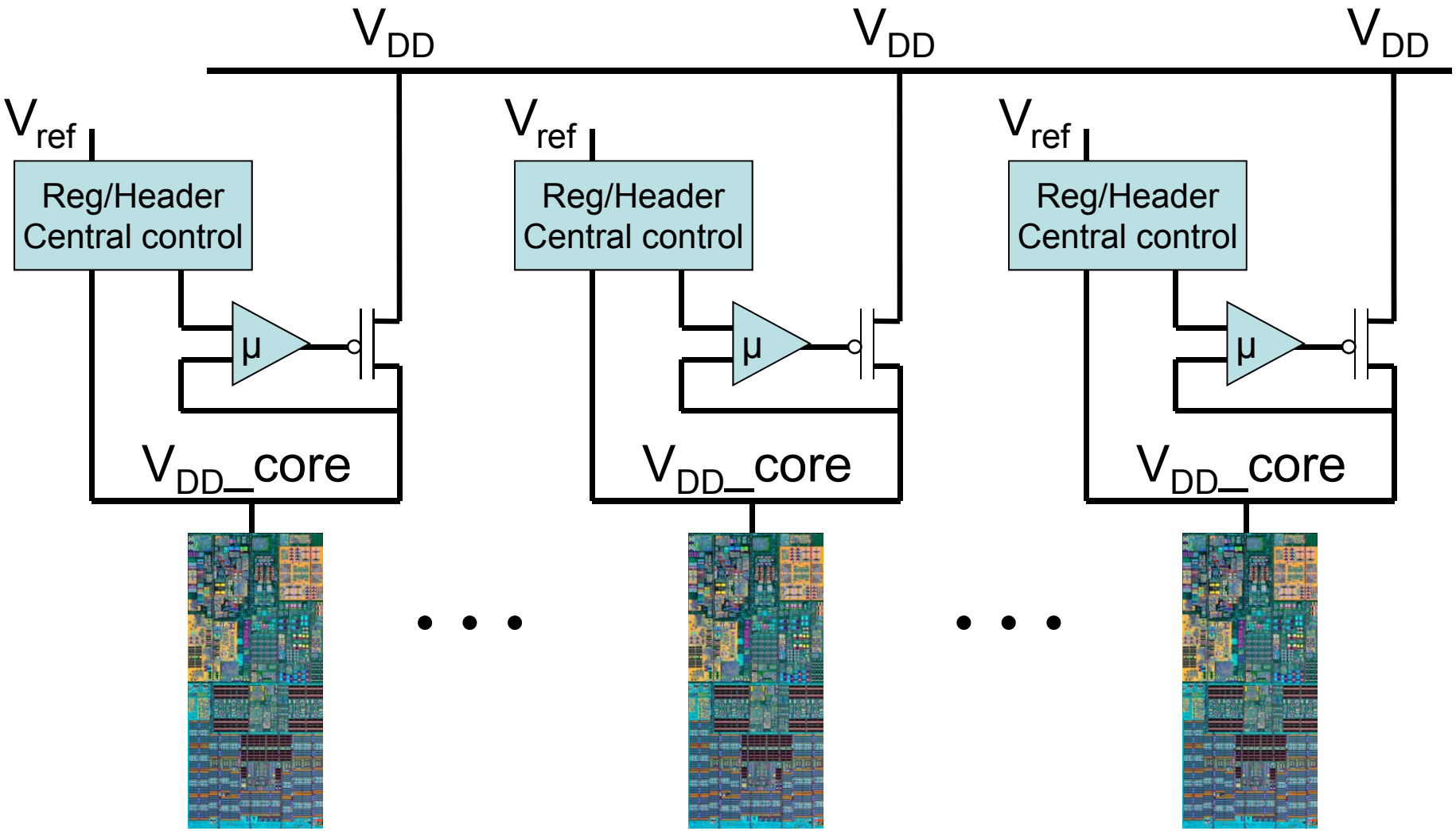
On-Die Per Core Power gating





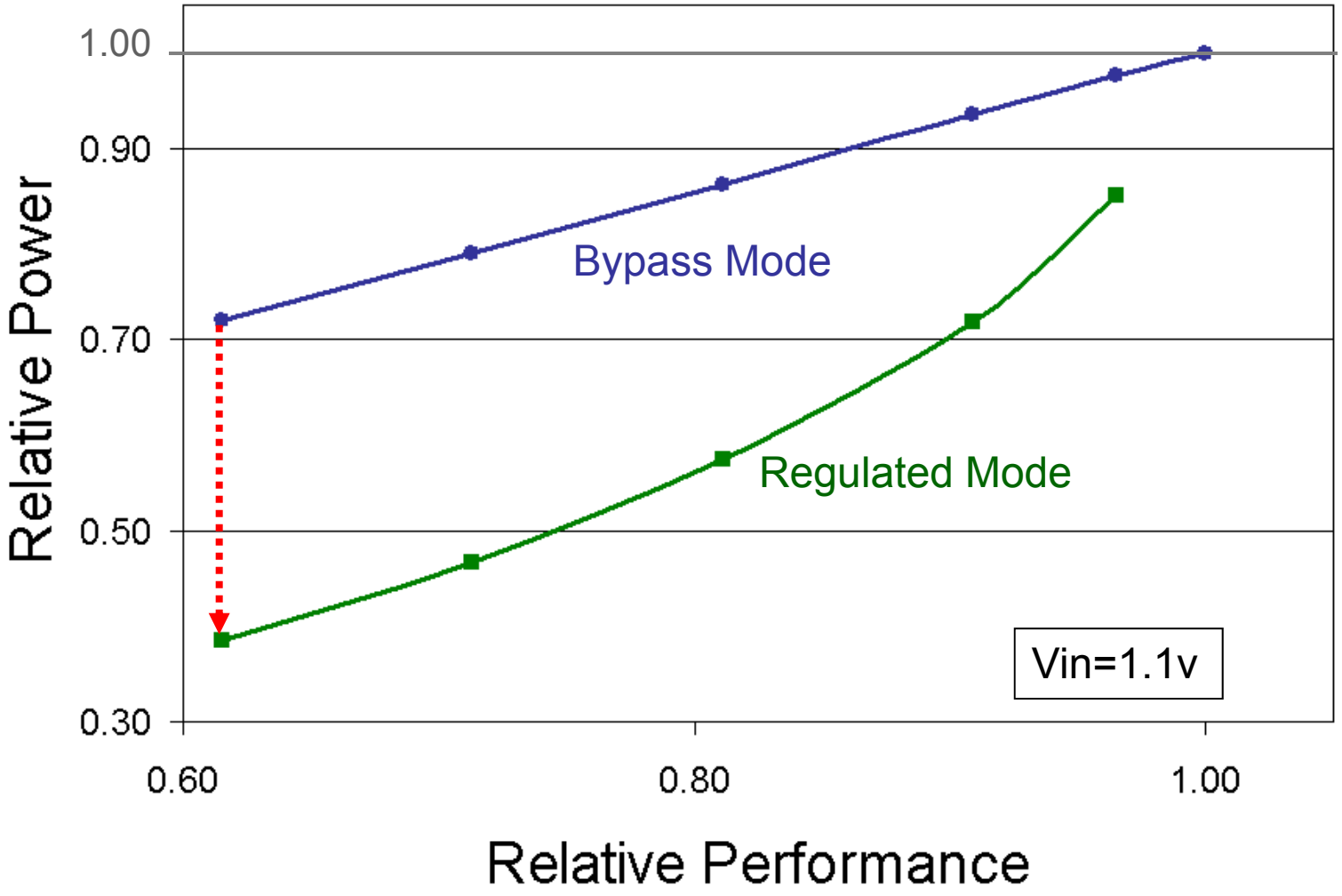
On-Die Per Core Voltage Regulation

- Each of the 12 core/cache partitions can adapt voltage to optimize power vs. performance demands



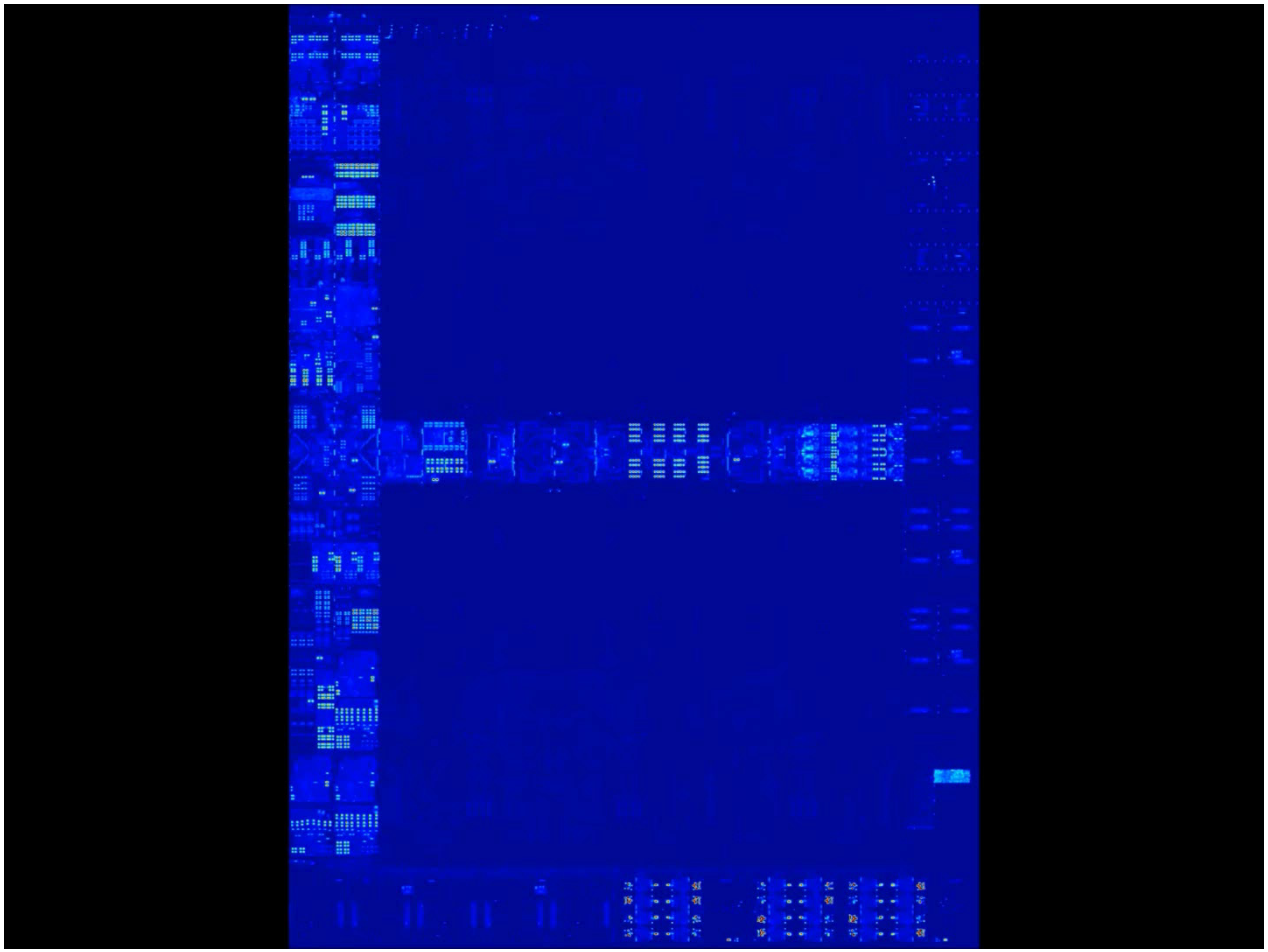
Voltage Regulation Benefit

- DVFS results vs DFS: ~33% power savings @ 62% freq



Per Core DVFS

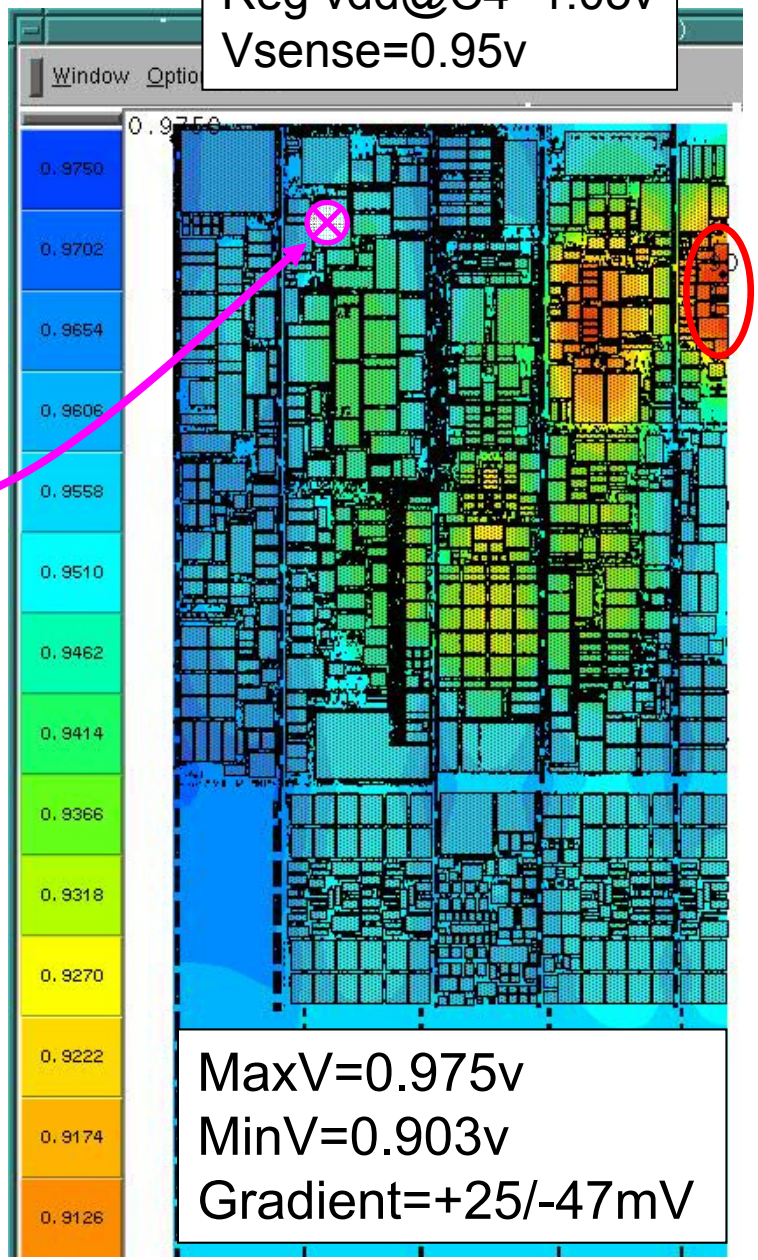
Freeze-frame video of ½ POWER8 per-core DVS using on-board VRMs



- Video creators: Franco Stellari, Al Weger, Chung-Ching Lin, and Peilin Song

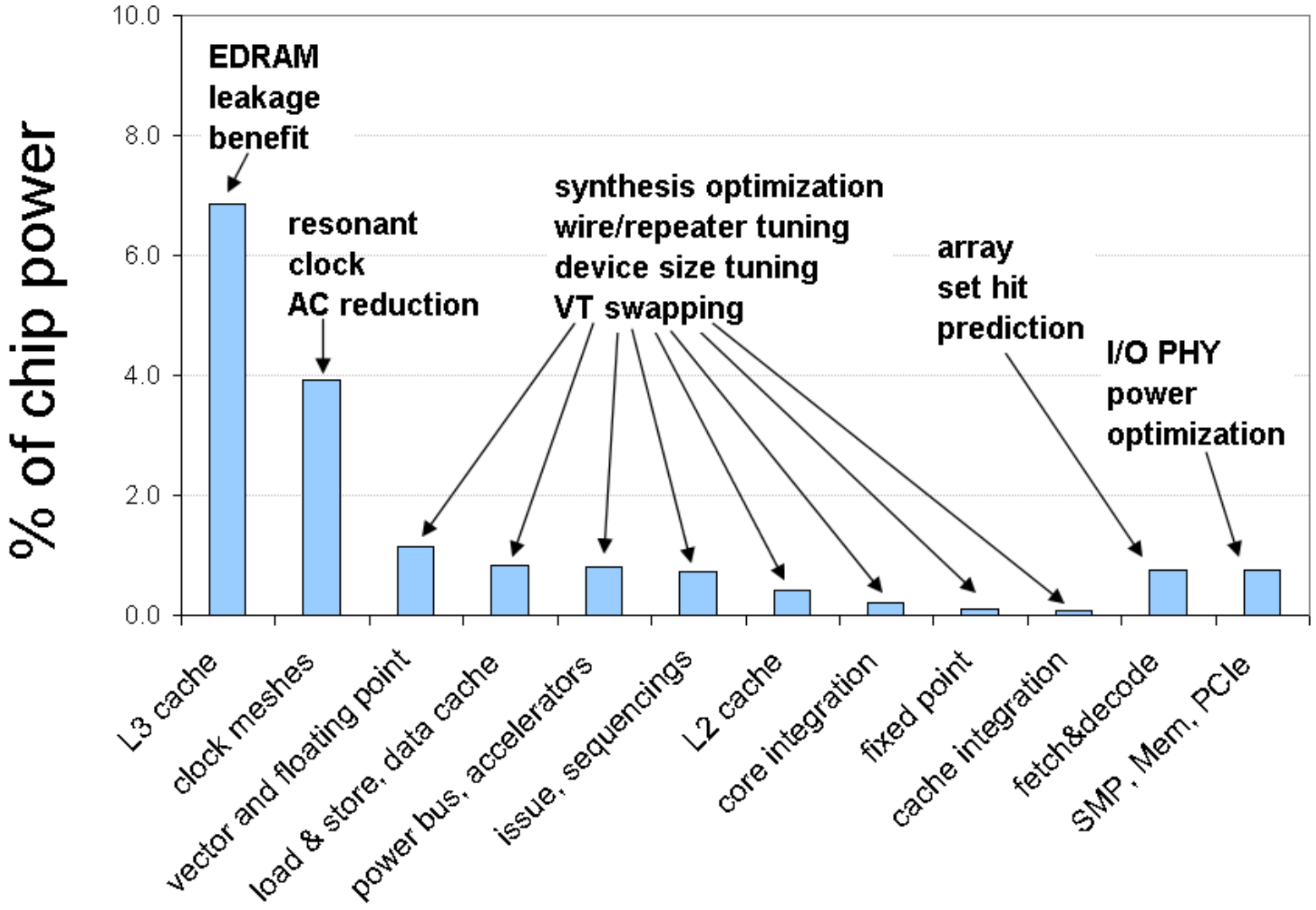
Power Grid Analysis

- **VRM sense point**
  - Will regulate supply such that sense point equals target Vsense (0.95v here)
- **Process**
  - Get worst-case load conditions, use DC to represent “steady-state” execution
  - Select a sense position, e.g.
  - Simulate region IR across workloads
  - Iterate to balance voltage gradients across core against Vsense goal
  - Physical header tuning can be used to dampen extreme differences from sense point
- **Demonstration simulation indicates the max deviations of +25mv/-47mv on right side of core, w.r.t Vsense**
- **Ultra-thick layers are key to low IR**



Power analysis & improvements

- Design effort = 17% savings of chip total power

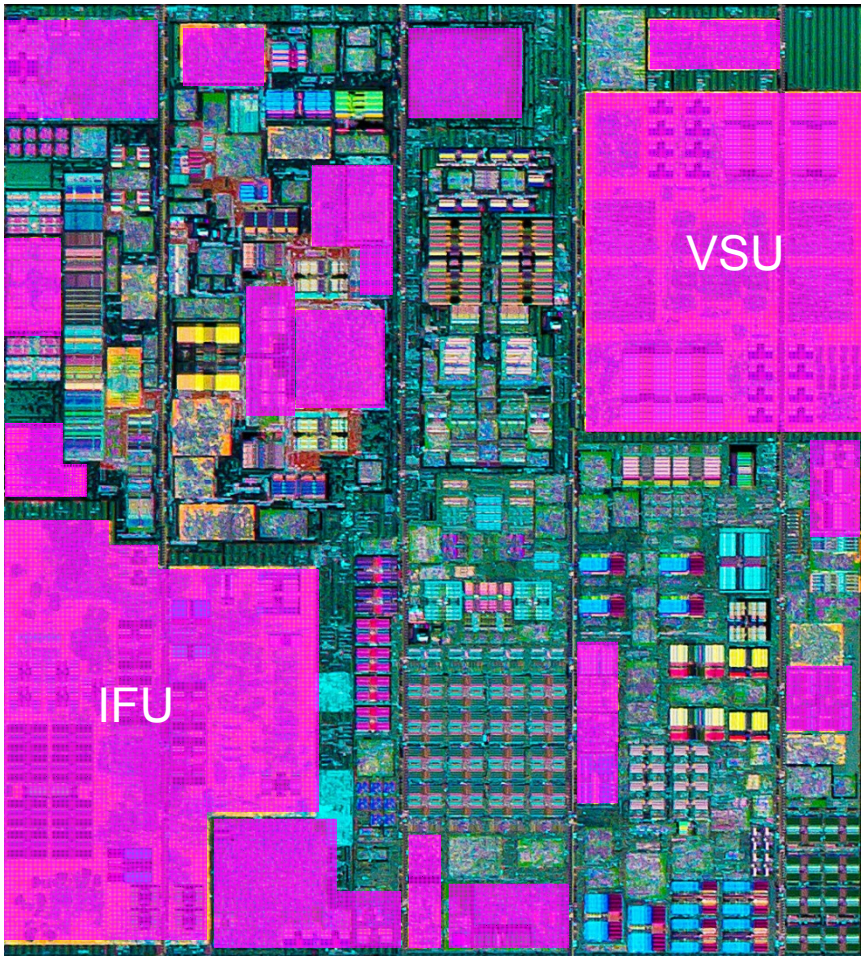


Outline

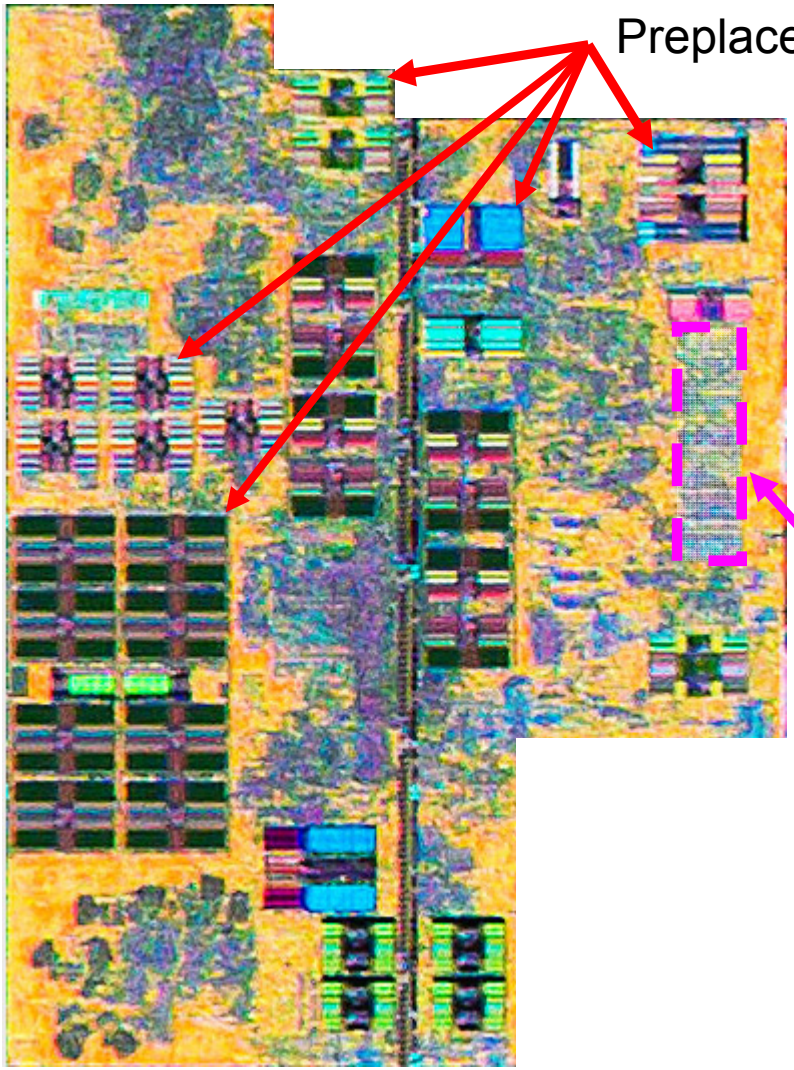
- Data optimized design
  - Technology
  - Highly threaded, wide execution core
  - High bandwidth nest
- Circuit optimizations
  - Arrays, Clocking, and IOs
- Power Management & Reduction
- Design methodology
- Lab data

Large Block Structured Synthesis

- Enhanced process which included:
  - Structured dataflow
  - Congestion-aware stdcell placement
  - Embedded “hard” IP (e.g. arrays, regfiles, complex custom cells)
  - Hierarchical embedded synthesis
- 30% fewer unique blocks vs. POWER7
- Improvements in block power and total design area
- Gate-level design TAT sign-off improvement of 3-10x

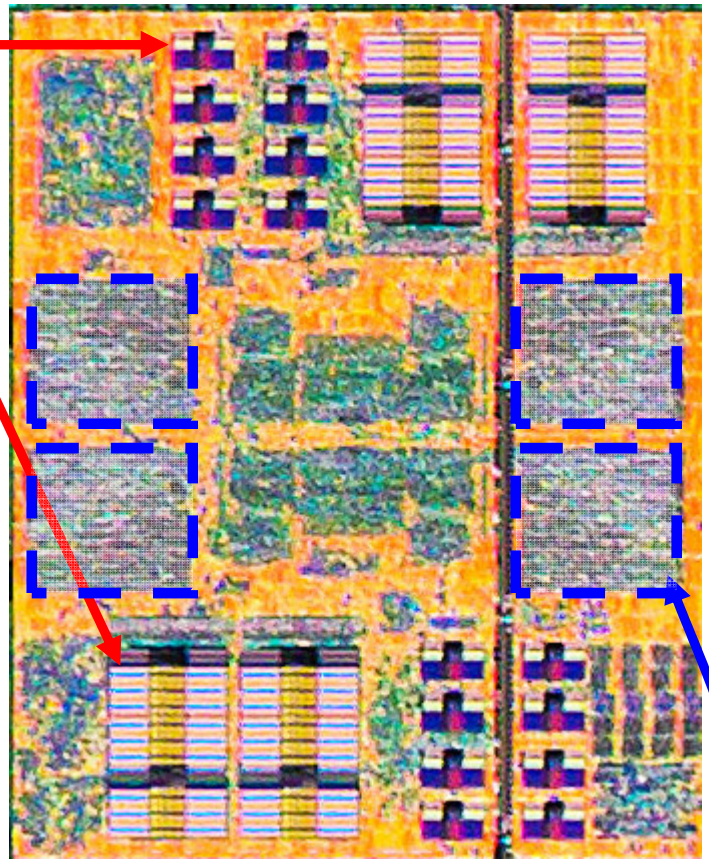


High Performance: IFU and VSU as LBSS



IFU: 580K gates, 628K nets

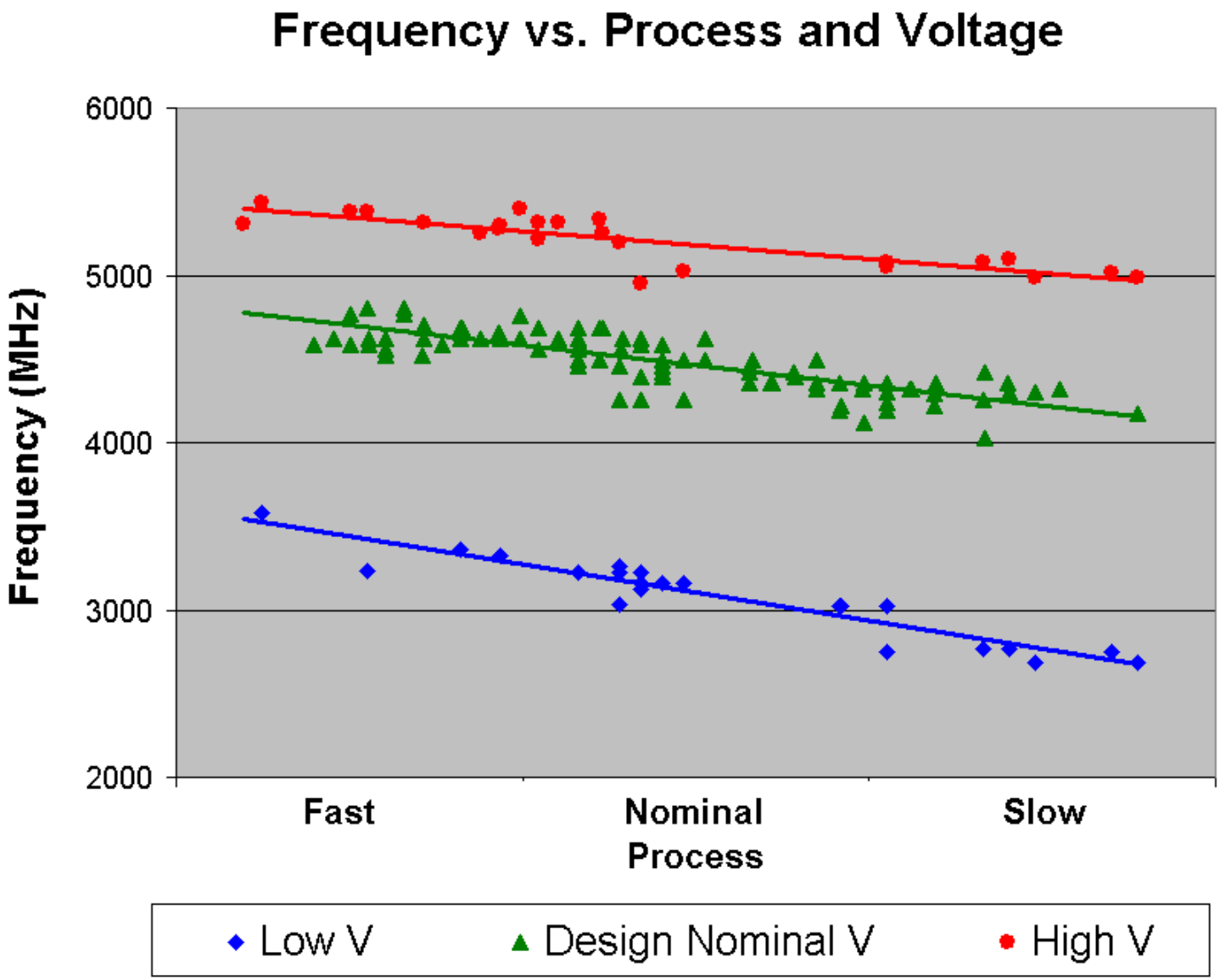
- 37 embedded array/register files



VSU: 697K gates, 723K nets

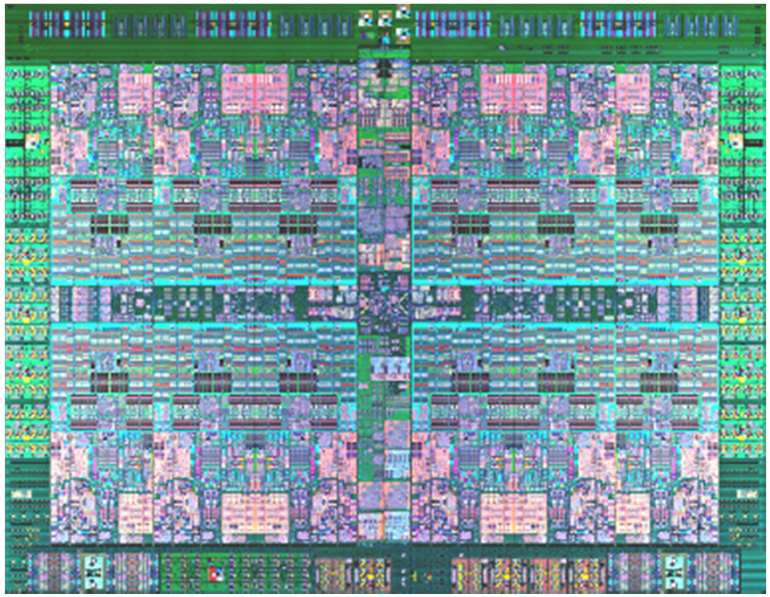
- 20 embedded array/register files
- hierarchical embedded synthesis

Characterization Data



Thanks!

POWERful!



POWER8™: A 12-Core Server-Class Processor in 22nm SOI with 7.6Tb/s Off-Chip Bandwidth

Distributed System of Digitally-Controlled Microregulators Enabling Per-Core DVFS for the POWER8™ Microprocessor

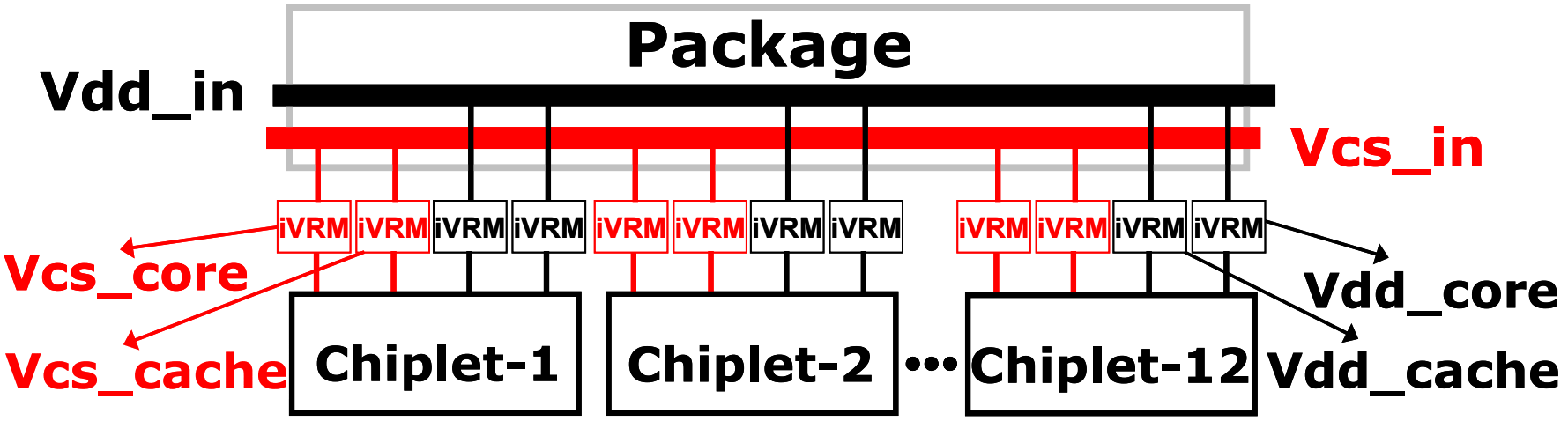
Zeynep Toprak-Deniz[1], Michael Sperling[2], John Bulzacchelli[1], Gregory Still[2], Ryan Kruse[2], Seongwon Kim[1], David Boerstler[2], Tilman Gloekler[2], Raphael Robertazzi[1], Kevin Stawiasz[1], Timothy Diemoz[2], George English[2], David Hui[2], Paul Muench[2], and Joshua Friedrich[2]

[1]IBM Research
[2]IBM Systems and Technology Group

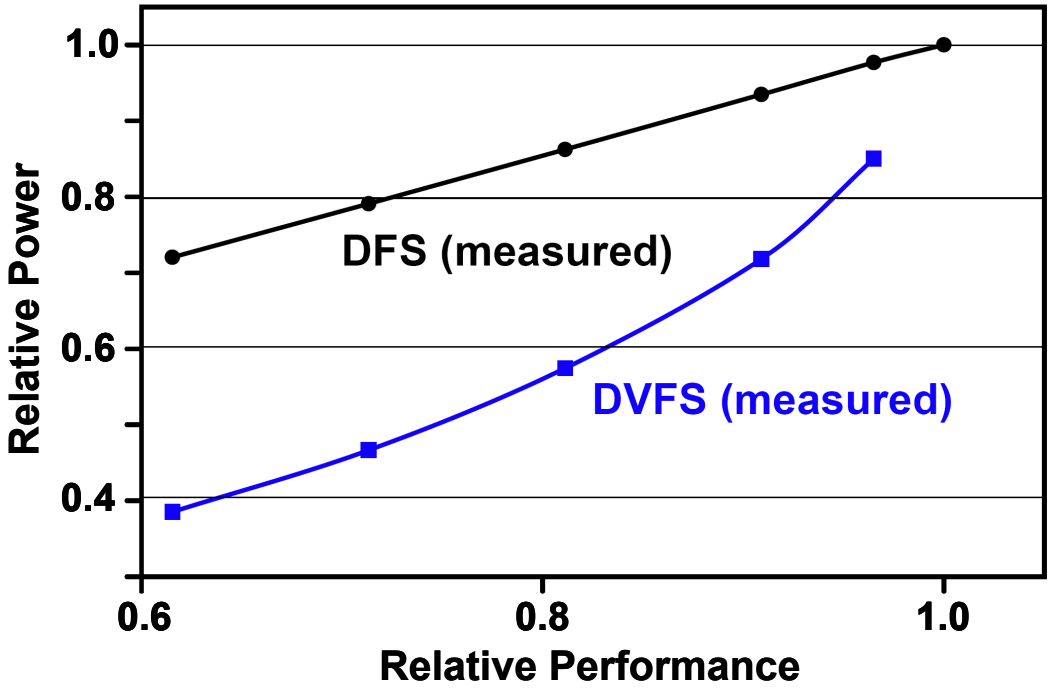
Outline

- Motivation
- Distributed architecture for integrated voltage regulator (iVRM)
- iVRM circuit designs
  - Voltage regulator controller (VREGC)
  - Microregulator (UREG)
- Experimental results
- Conclusions

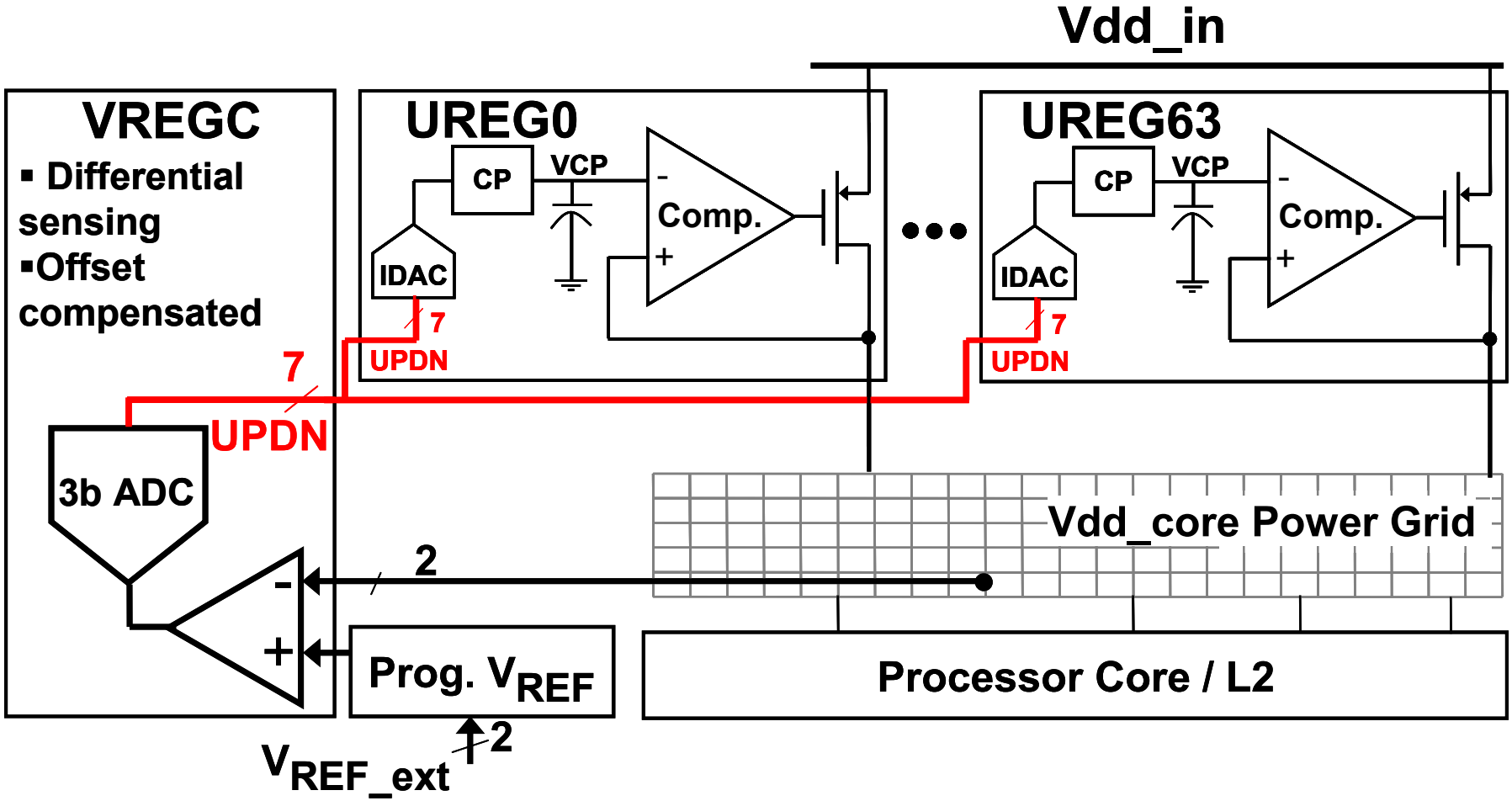
Motivation



- Per-core DVFS for performance optimization of power-constrained multi-core processor
- 4 regulated domains per chiplet (48 total)



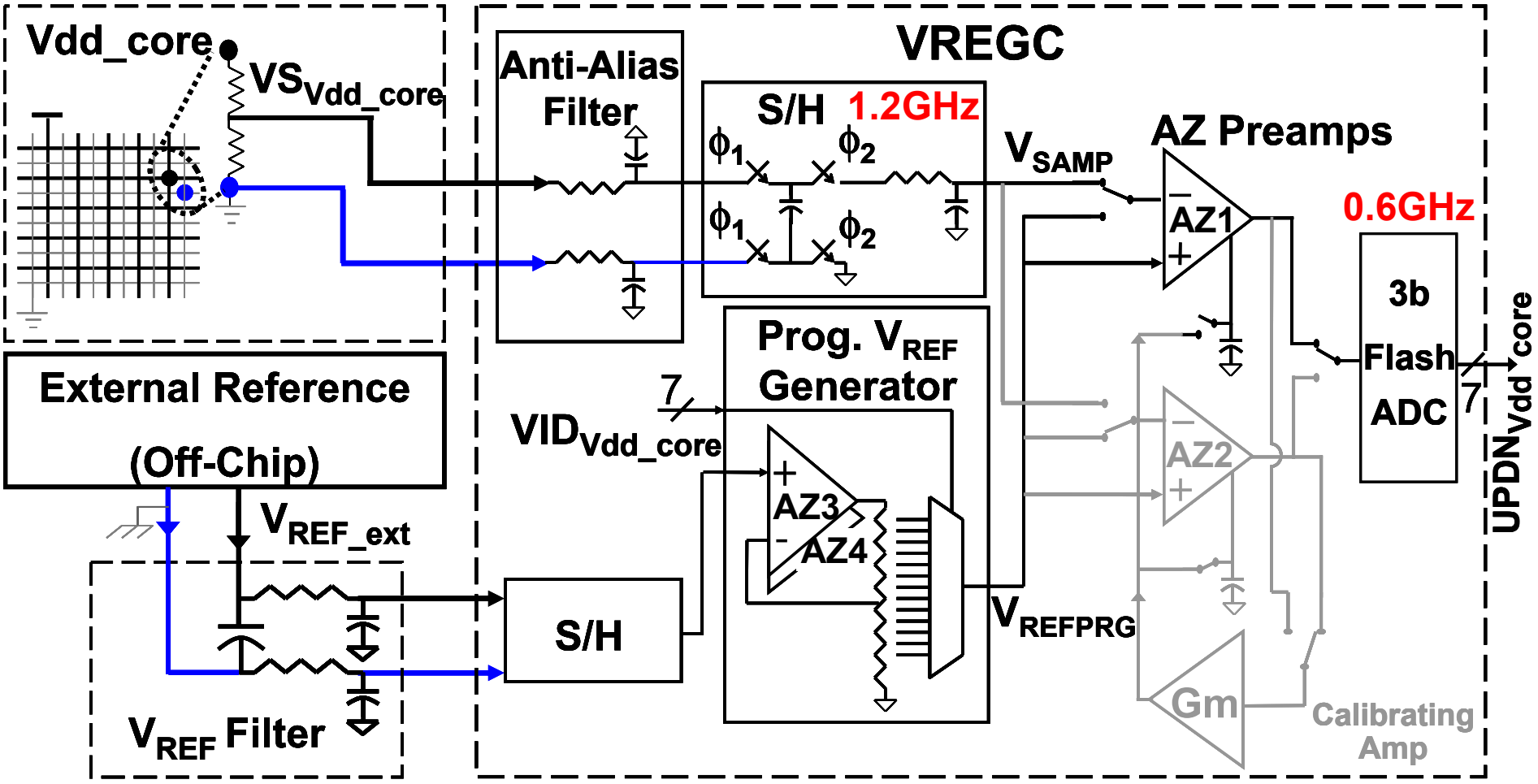
Dual Loop Distributed Regulator Architecture [3]



- External reference voltage used for high precision

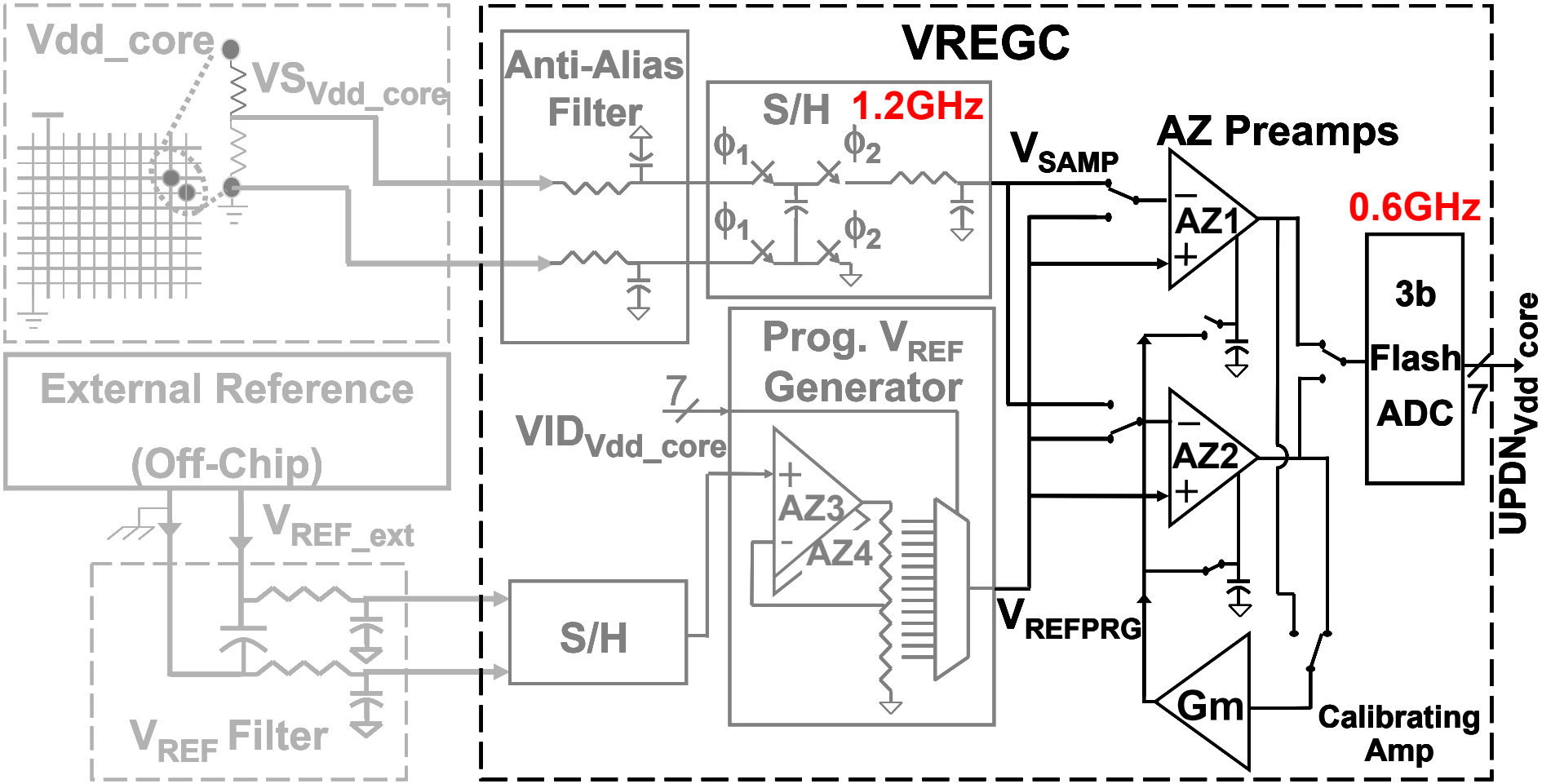
[3] J. F. Bulzacchelli et al., *JSSC*, April. 2012

Voltage Regulator Controller (VREGC)



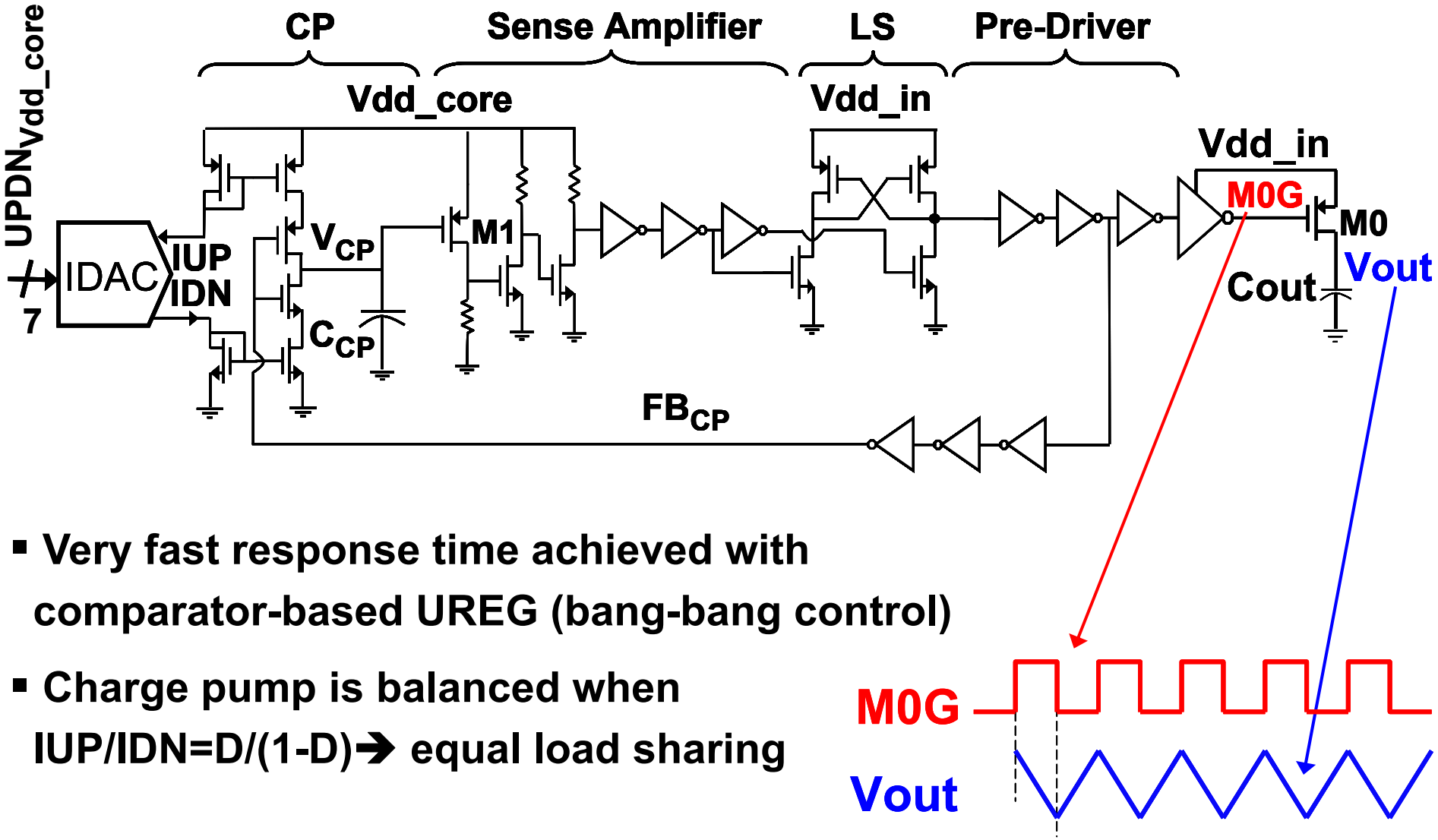
- Regulated voltage is filtered before sampling to avoid aliasing of high-frequency noise/ripple
- S/H also accomplishes differential to single ended conversion

Voltage Regulator Controller (VREGC)



- Ping-pong architecture for auto-zeroing preamplifiers

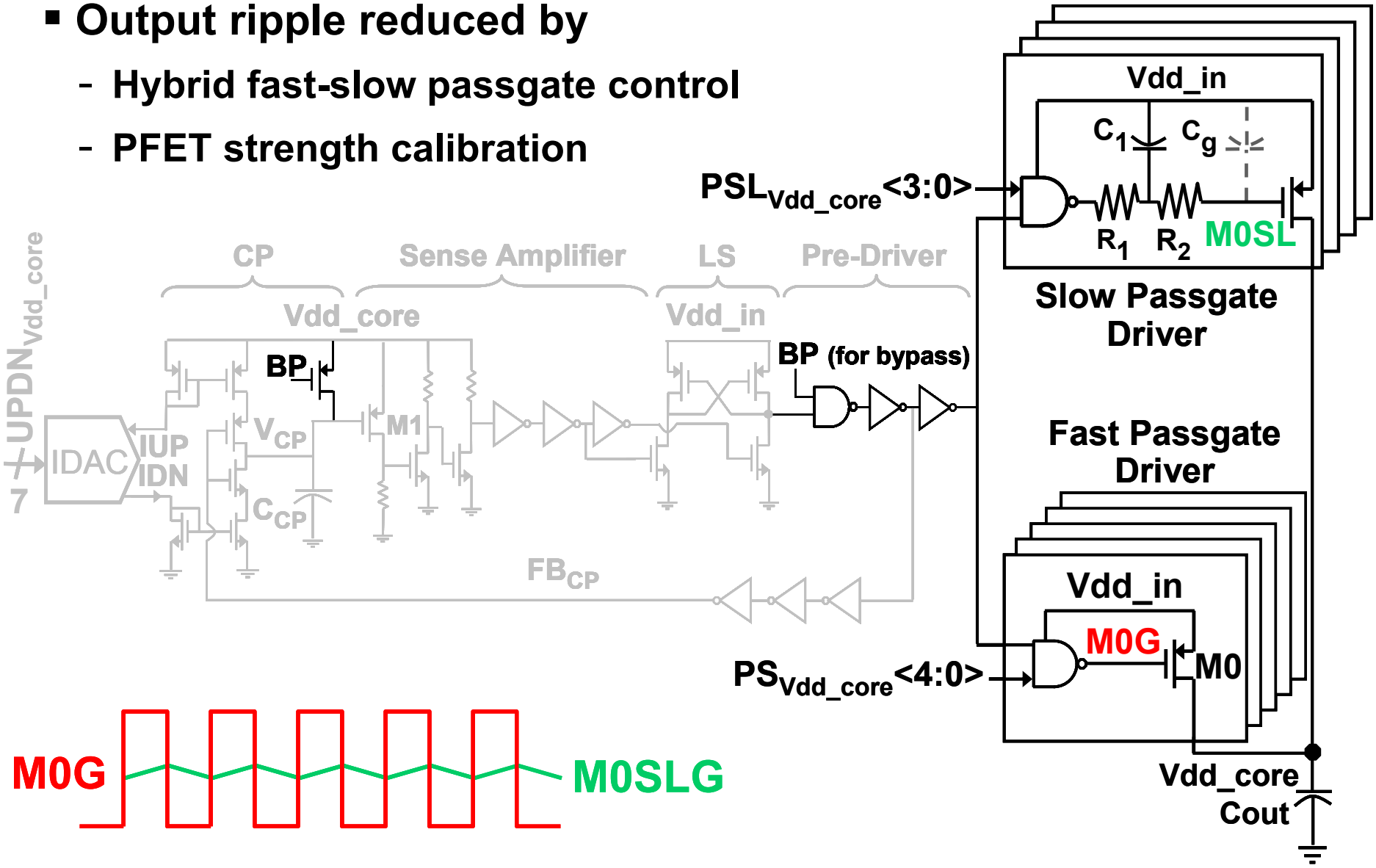
Basic Microregulator (UREG) Structure



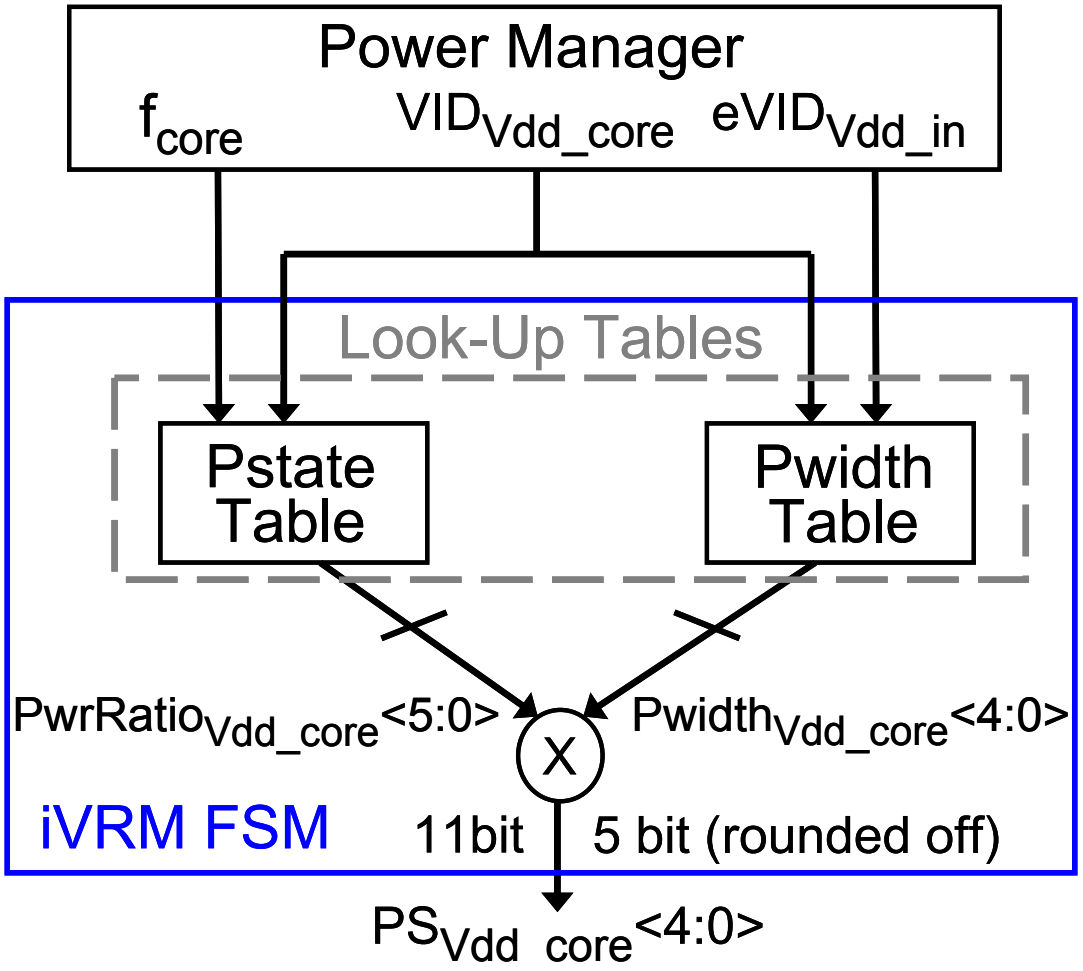
- Very fast response time achieved with comparator-based UREG (bang-bang control)
- Charge pump is balanced when IUP/IDN=D/(1-D)➔ equal load sharing

Ripple Reduction Techniques

- Output ripple reduced by
  - Hybrid fast-slow passgate control
  - PFET strength calibration



Predictive PFET Strength Calibration

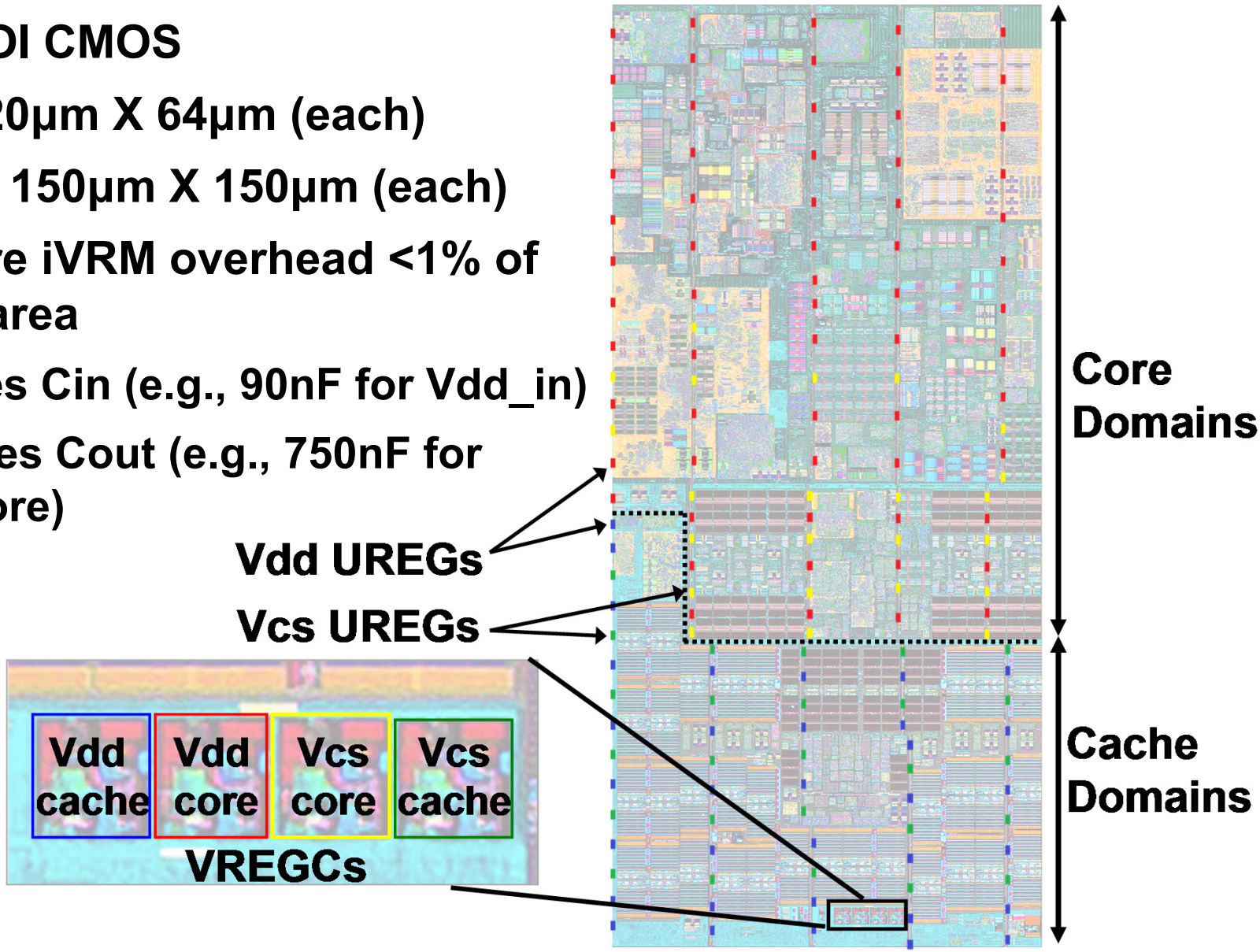


- **Pstate Table: stores max. load current information as function of output voltage and f_{core}**
- **Pwidth Table: records passgate width needed to support given drain current as function of input and output voltages**

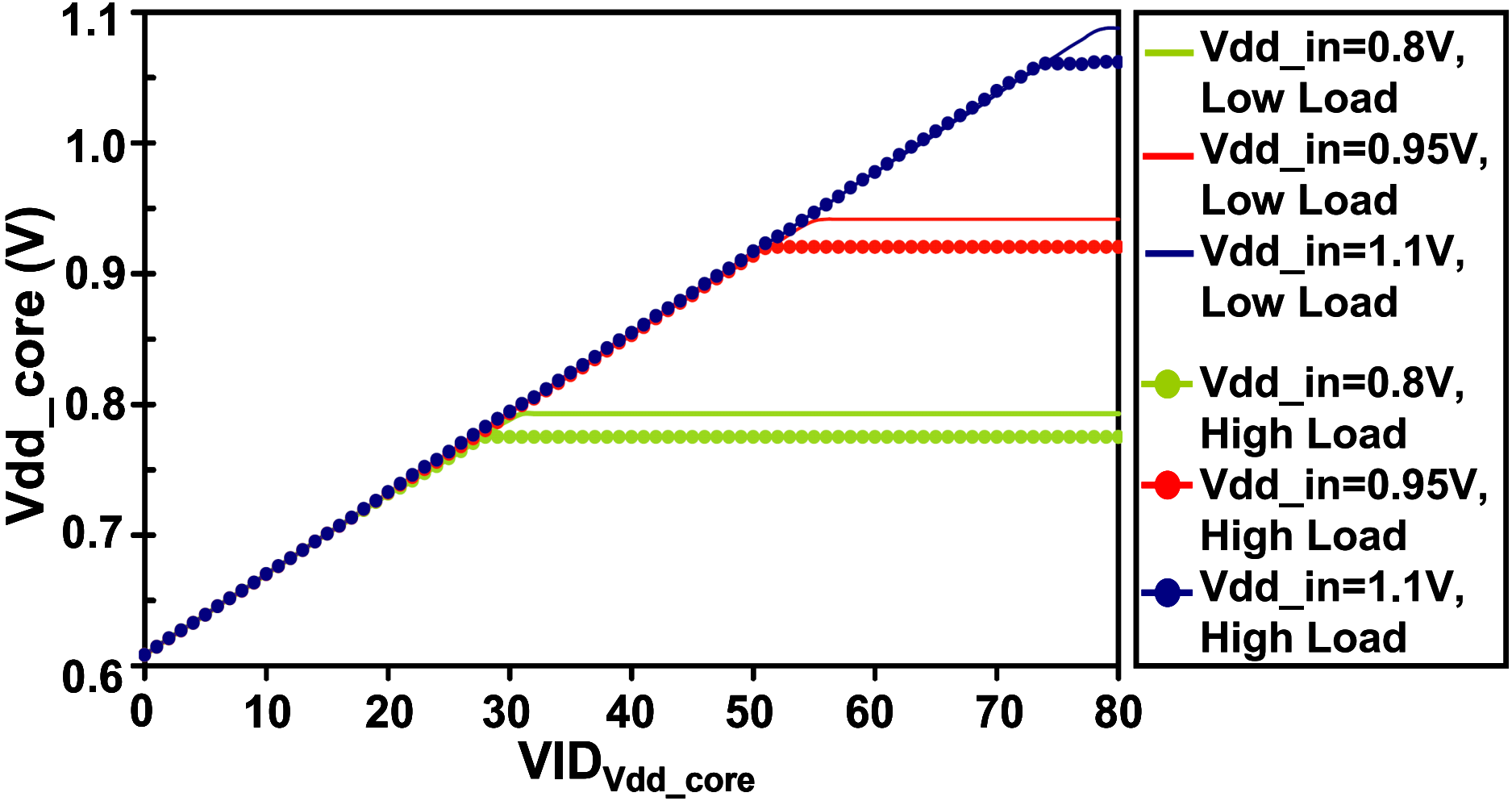
PFETStrength = PstateTable • PwidthTable

Micrograph of POWER8™ Chiplet

- 22nm SOI CMOS
- UREG: 20µm X 64µm (each)
- VREGC: 150µm X 150µm (each)
- Vdd_core iVRM overhead <1% of chiplet area
  - Includes Cin (e.g., 90nF for Vdd_in)
  - Excludes Cout (e.g., 750nF for Vdd_core)

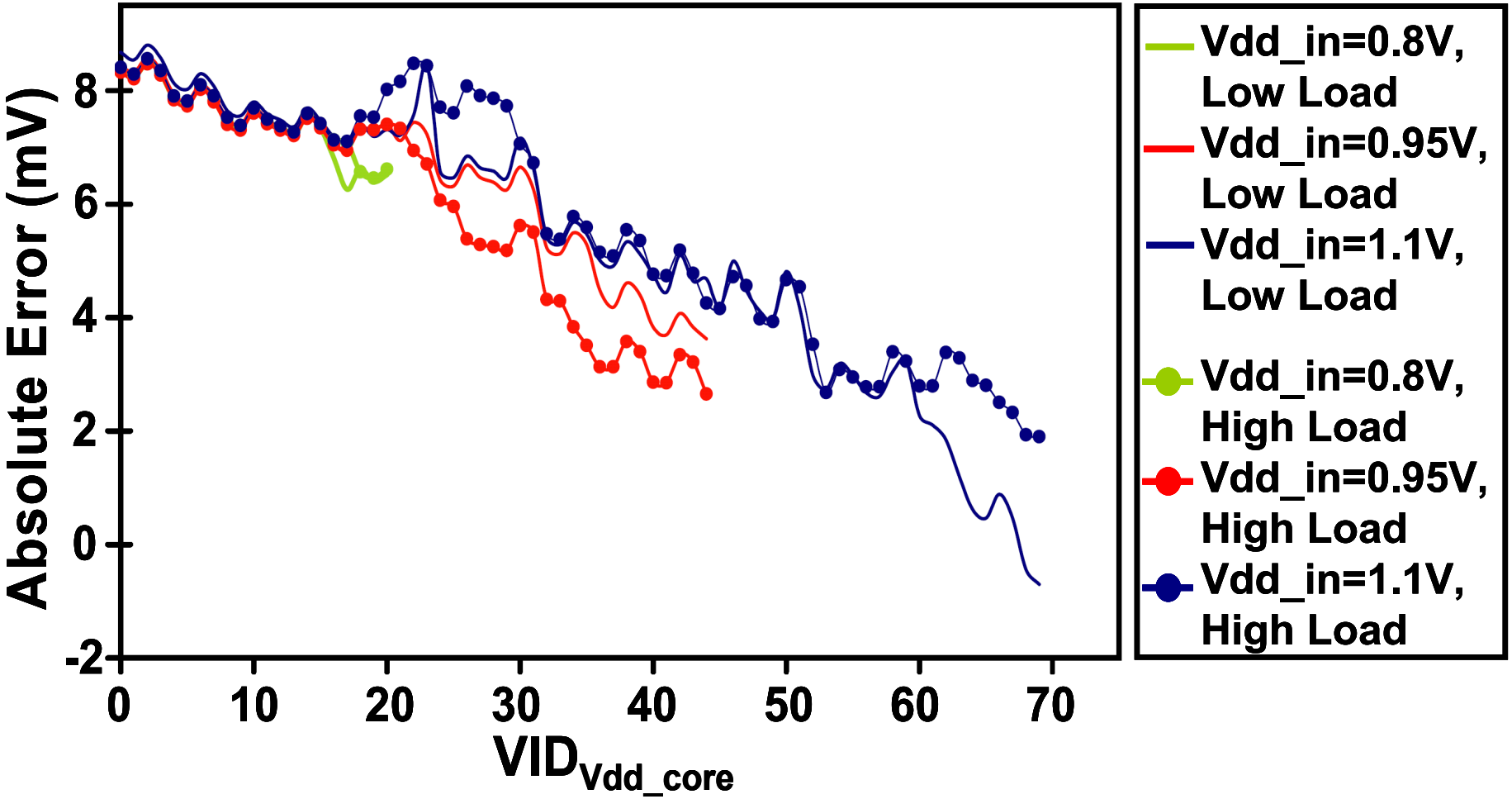


Measured DC Voltage versus VID



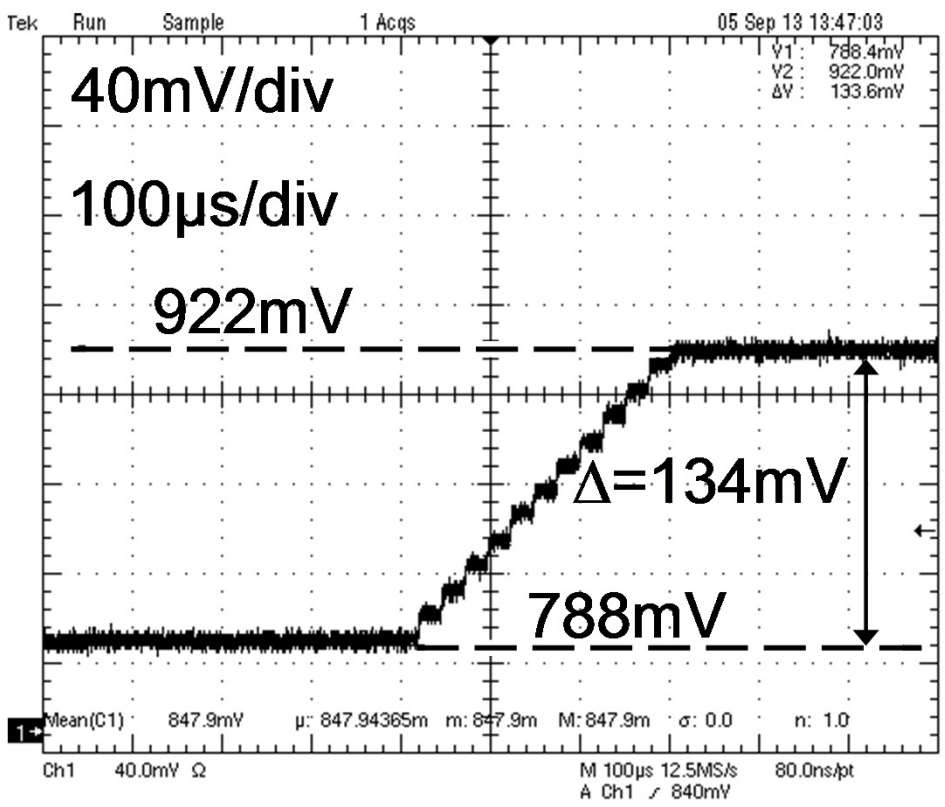
- Dropout voltage < 50mV

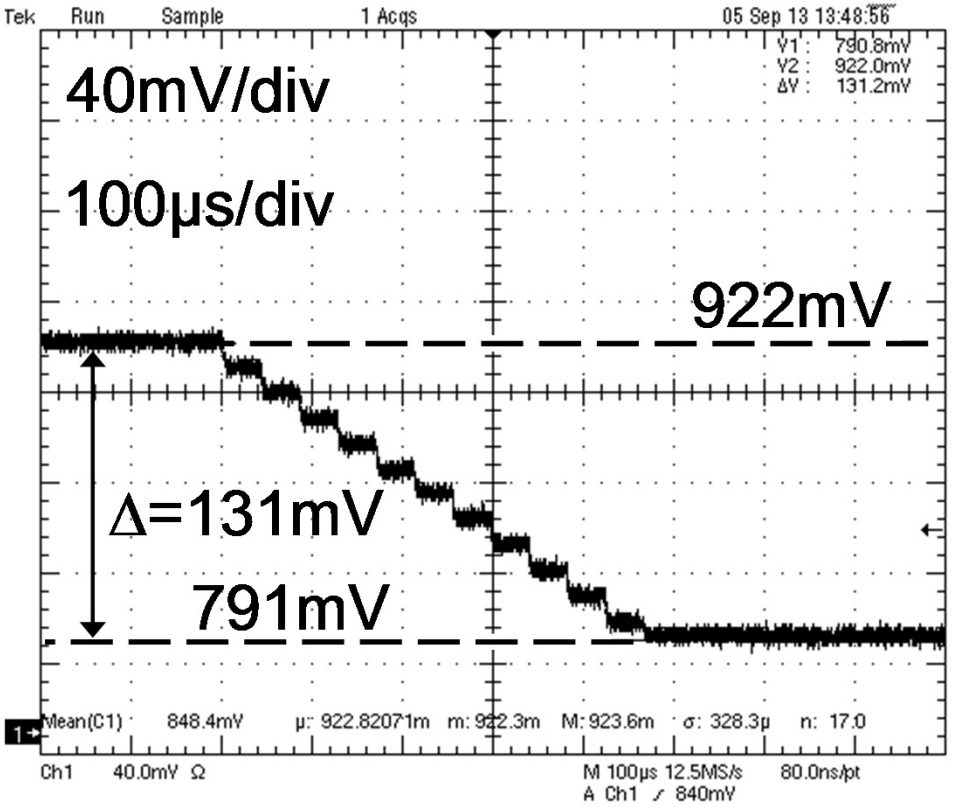
Measured Absolute Voltage Error



- Absolute voltage error < 9mV
- Load regulation error < 3mV
- Variation with Vdd_in < 5mV

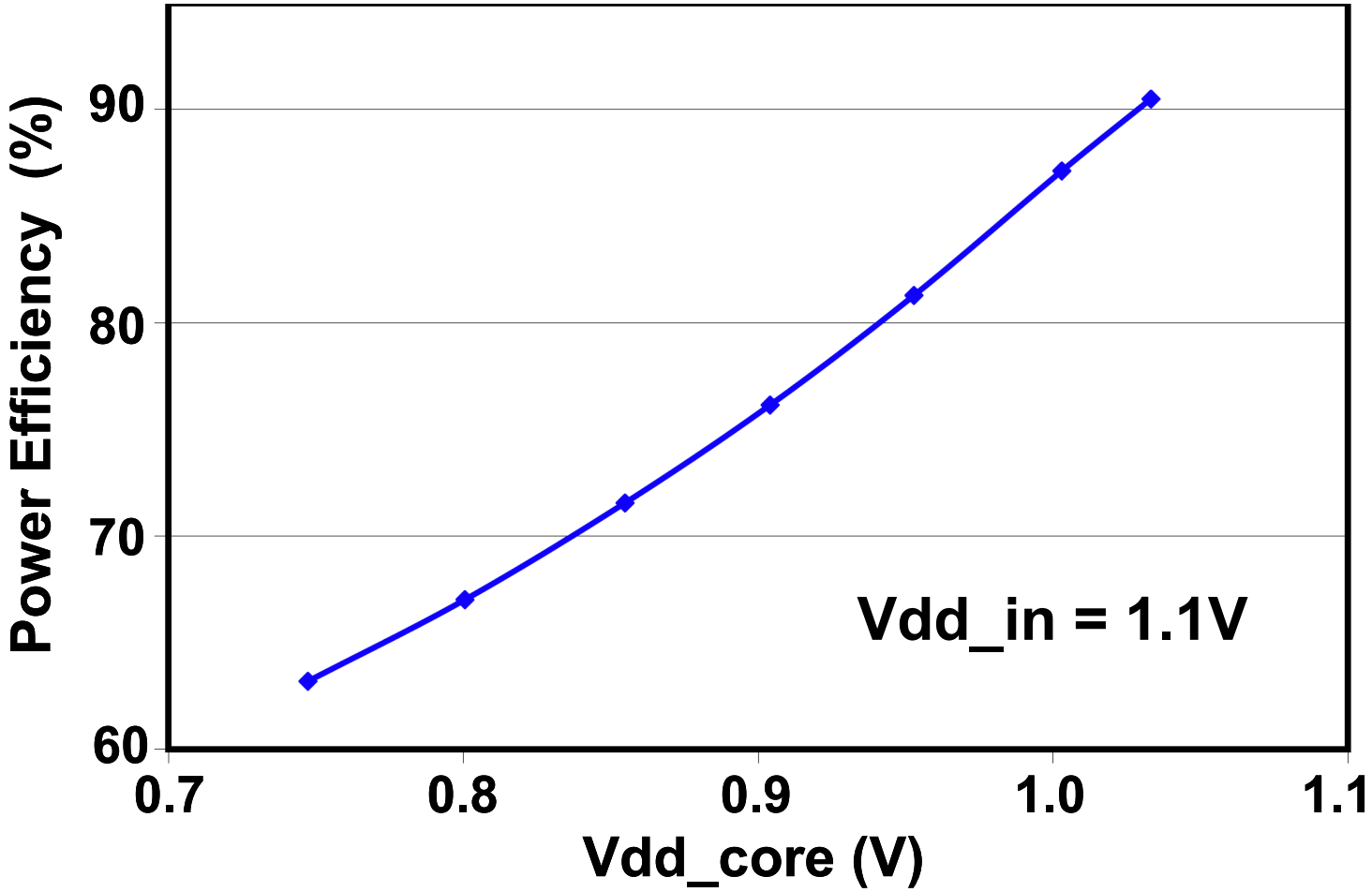
Output Voltage Stepping





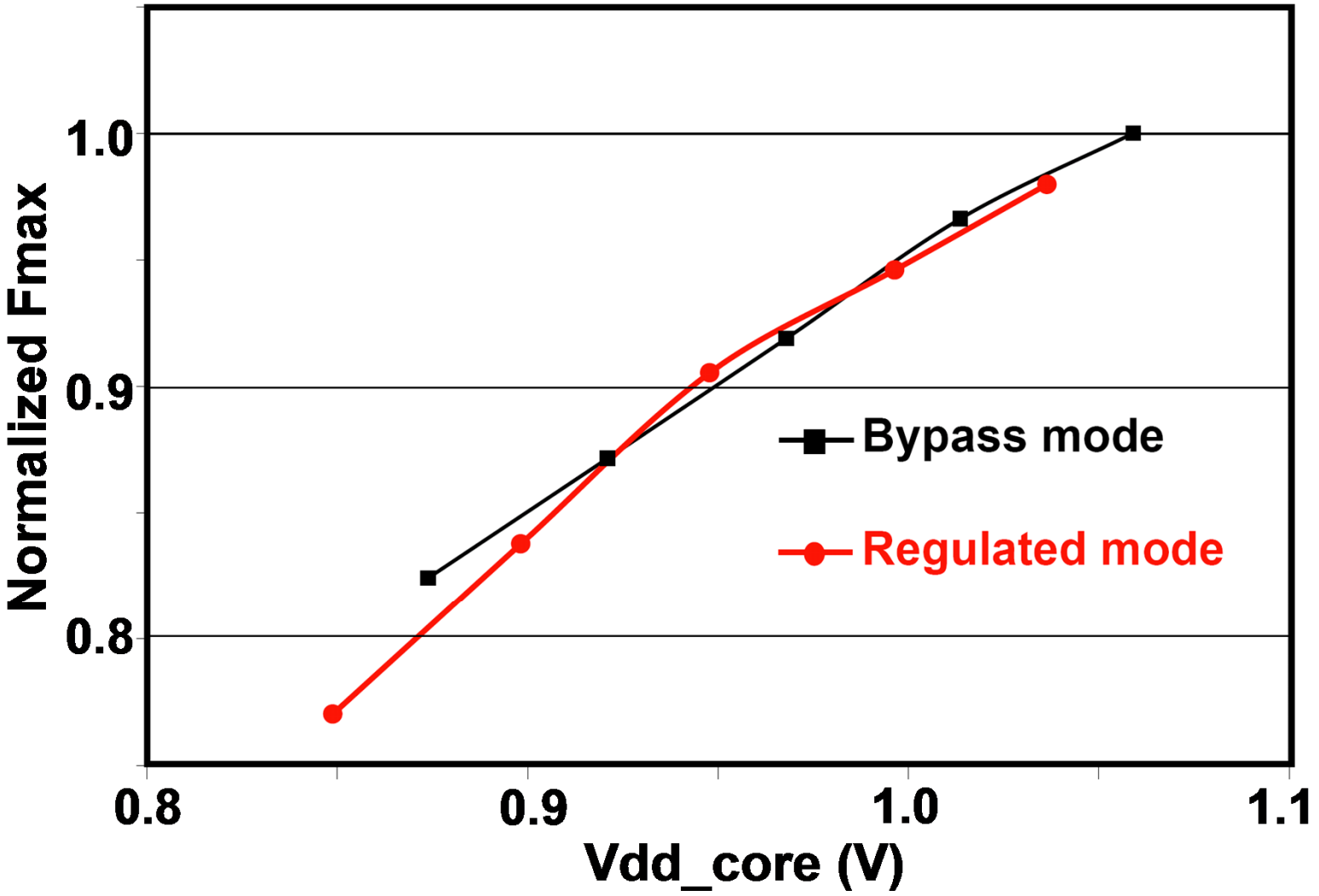
- **Small steps (e.g., 12.5mV) ensure tracking between domains**

Power Efficiency (under High Load)



- Peak power efficiency of 90.5% at Vdd_core =1.03V
  - Load current = 12.3A
  - iVRM power density = 36W/mm^2

Normalized Fmax versus Vdd_core



- Fmax normalized to its value at Vdd_core =1.05V in bypass mode

Conclusions

- iVRMs enable per-core DVFS for POWER8™ processor
  - First reported integrated voltage regulator system with no off-chip components for server-class processor
- Key features
  - Dual loop system of distributed UREGs
    - total of 1764 UREGs per chip
  - Digital distribution of UPDN codes
  - Predictive PFET strength calibration based on look-up tables
- Measured performance
  - Peak power efficiency of 90.5%
  - Power density of 36W/mm^2
    - at least 3.5X higher than that of inductor-based or SC converters

Acknowledgments

- Members of IBM Research and IBM Systems and Technology Group for essential design, test, and managerial support

- Specific IBM colleagues:

  L. Acevedo and A. Wu for critical verification work

Wide-Frequency-Range Resonant Clock with On-the-Fly Mode Changing for the POWER8™ Microprocessor

Phillip Restle[1], David Shan[2], David Hogenmiller[2],
Yong Kim[2], Alan Drake[3], Jason Hibbeler[4], Thomas Bucelot[1],
Gregory Still[5], Keith Jenkins[1], Joshua Friedrich[2]

IBM Research:

1-Yorktown Heights, NY

3-Austin, TX

IBM Systems and Technology Group:

2-Austin, TX

4-Williston, VT

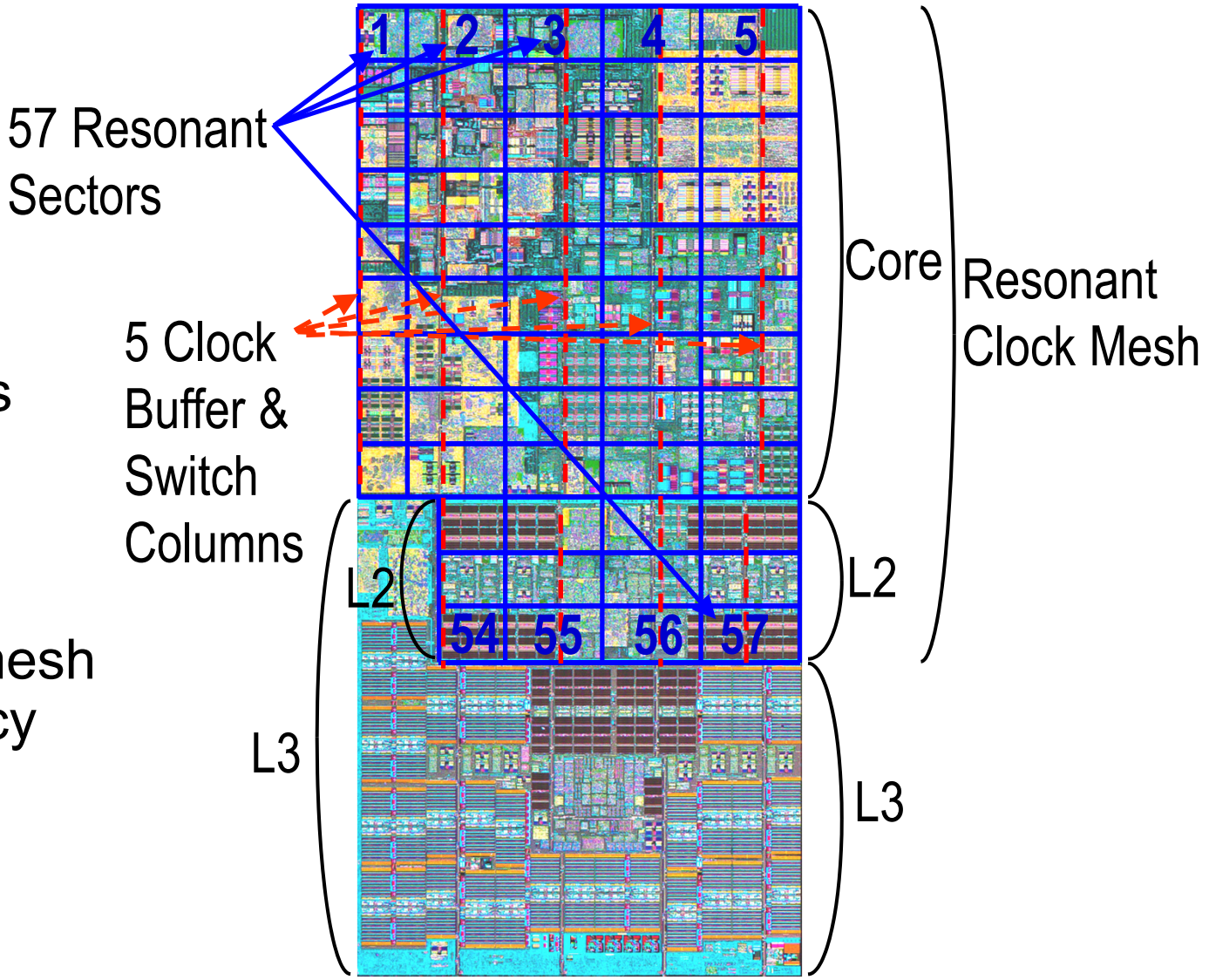
5-Raleigh, NC

Outline

- POWER8™ Chiplet Clock Meshes
- Resonant circuits
  - Drivers, Switches, Capacitors, Inductors
- New Features:
  - Two resonant frequencies
  - On-The-Fly Mode Changing
- Hardware results
- Conclude

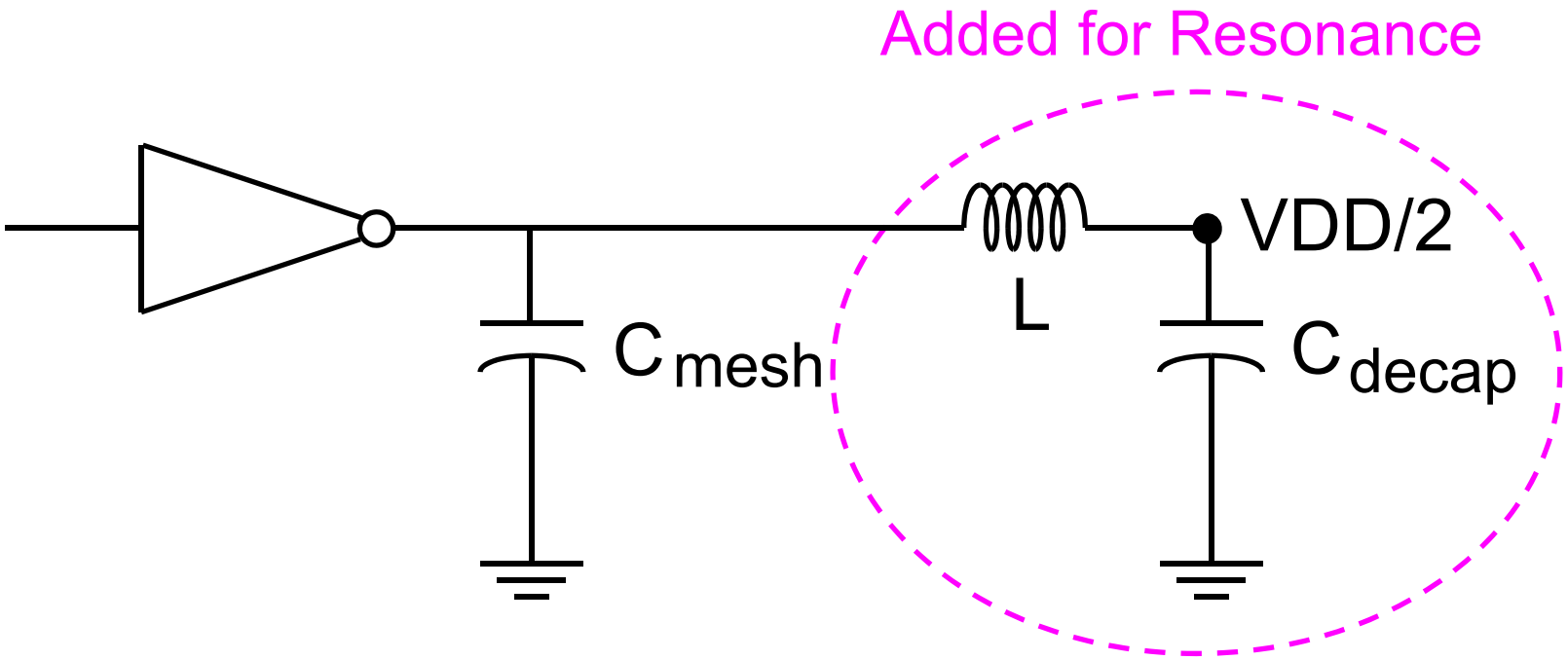
Chiplet Die Photo

- Core & L2 uses resonant clock with 57 sectors
- The L3 clock mesh is a ½ frequency non-resonant clock domain



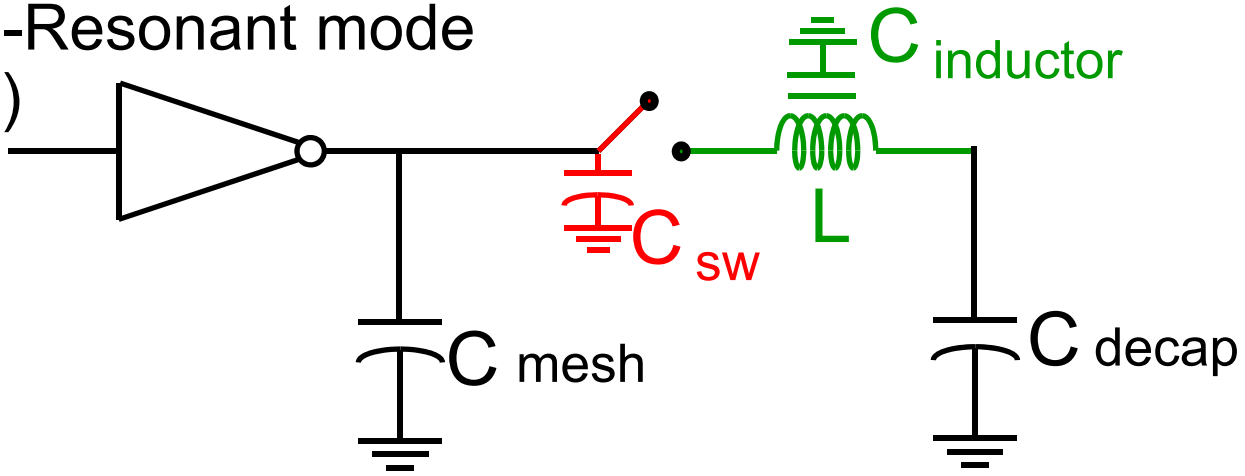
Simplest Form

$$F_{resonant} = \frac{1}{2\pi\sqrt{LC_{mesh}}} \qquad (C_{decap} >> C_{mesh})$$

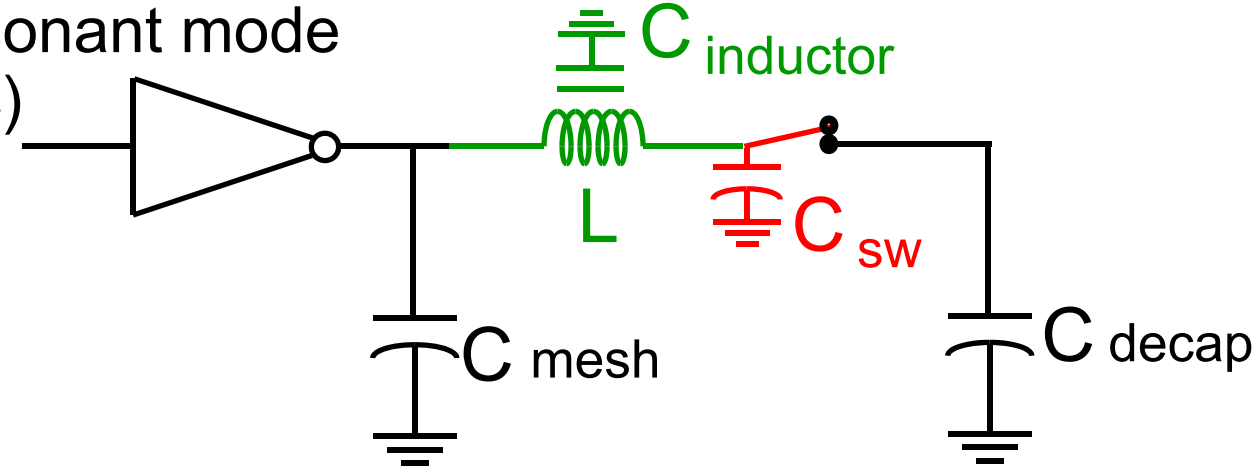


Mode Switch Choices

- Optimum for Non-Resonant mode ($C_{inductor}$ removed)

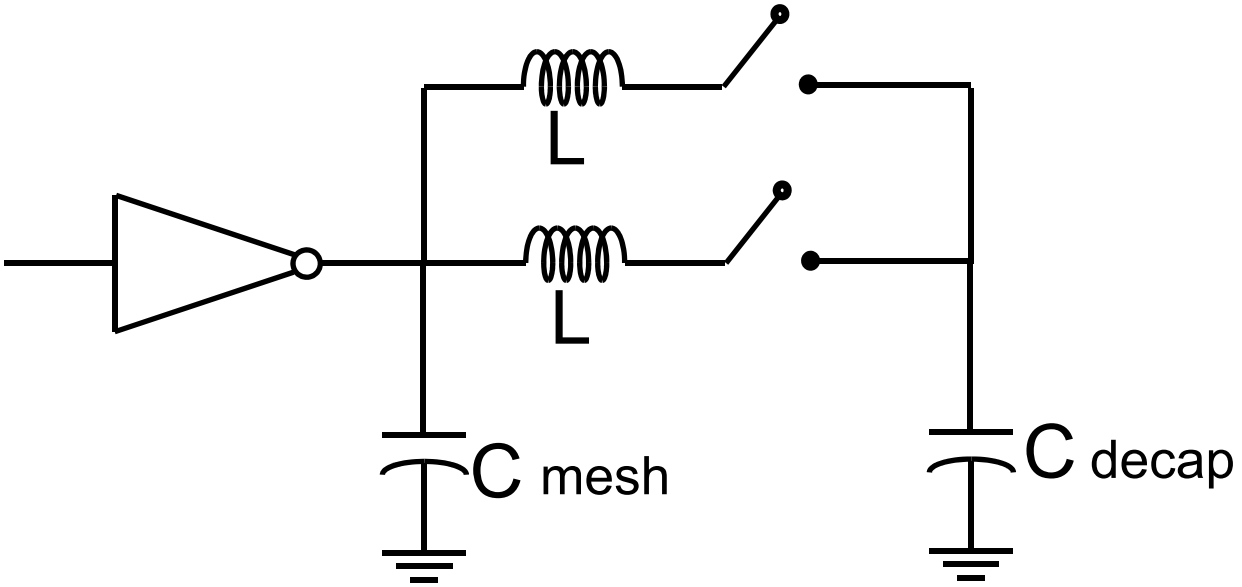


- Optimum for Resonant mode (C_{sw} not parasitic)



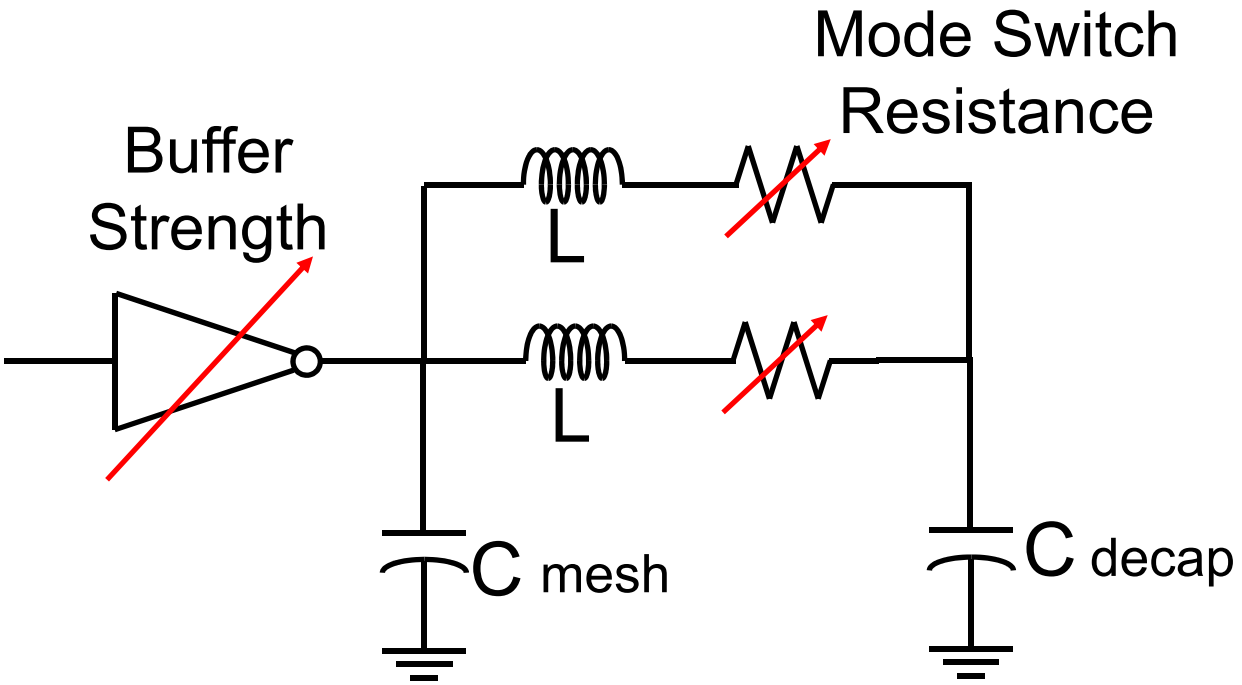
Two Inductors needed for Wide-Frequency-Range

- To save power from 2.5 to >5 GHz a second inductor and switch were added
- NR_{clk} Non-Resonant mode - both switches open
- LF_{clk} Low-Frequency resonant mode – one switch closed
- HF_{clk} High-Frequency resonant mode - both switches closed



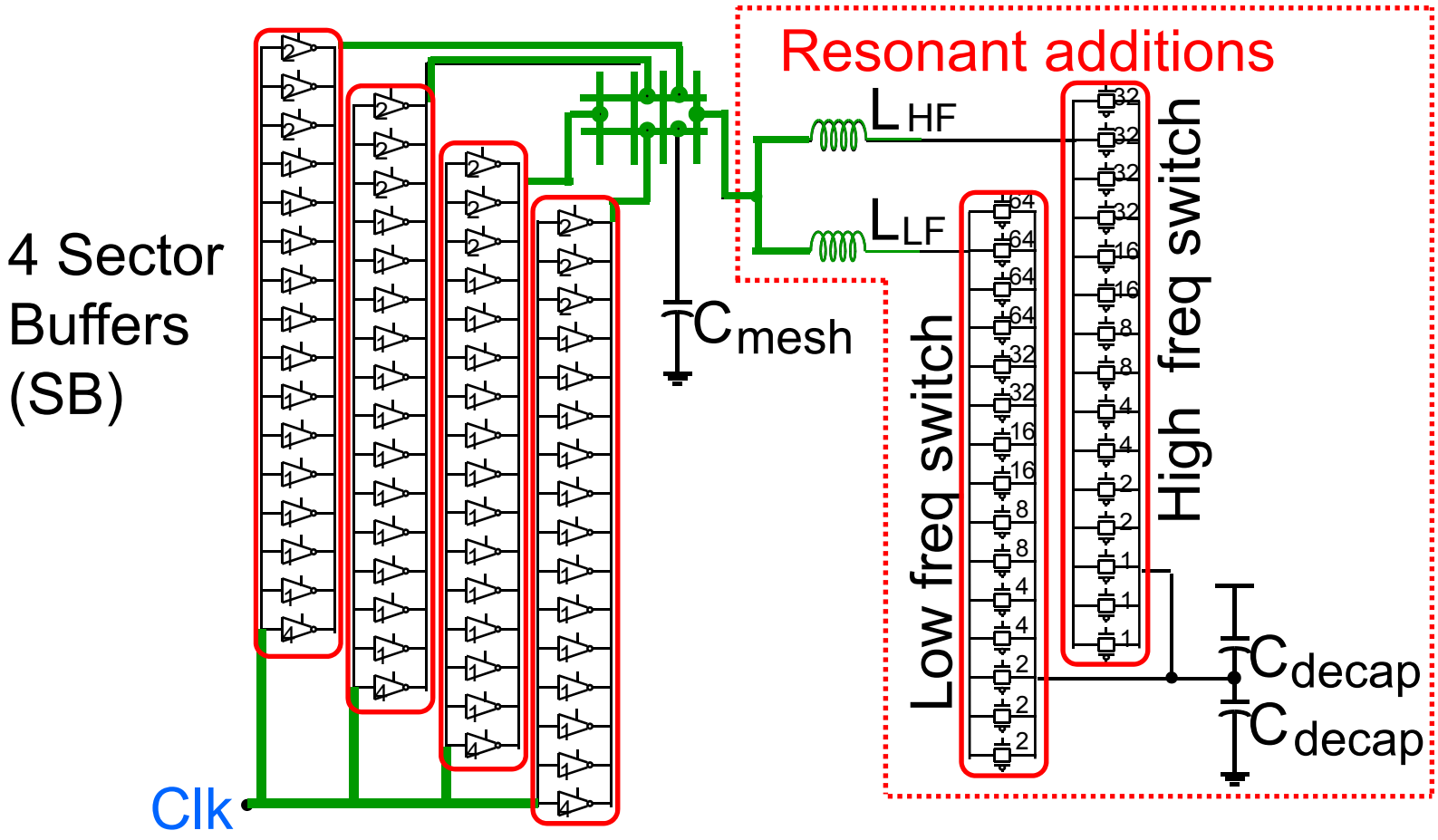
On-The-Fly Mode Changing

- Gently change modes with no impact on performance
  - Requires fine dynamic control of buffer strength and switch resistance
  - Controlled by OCC (on chip controller)



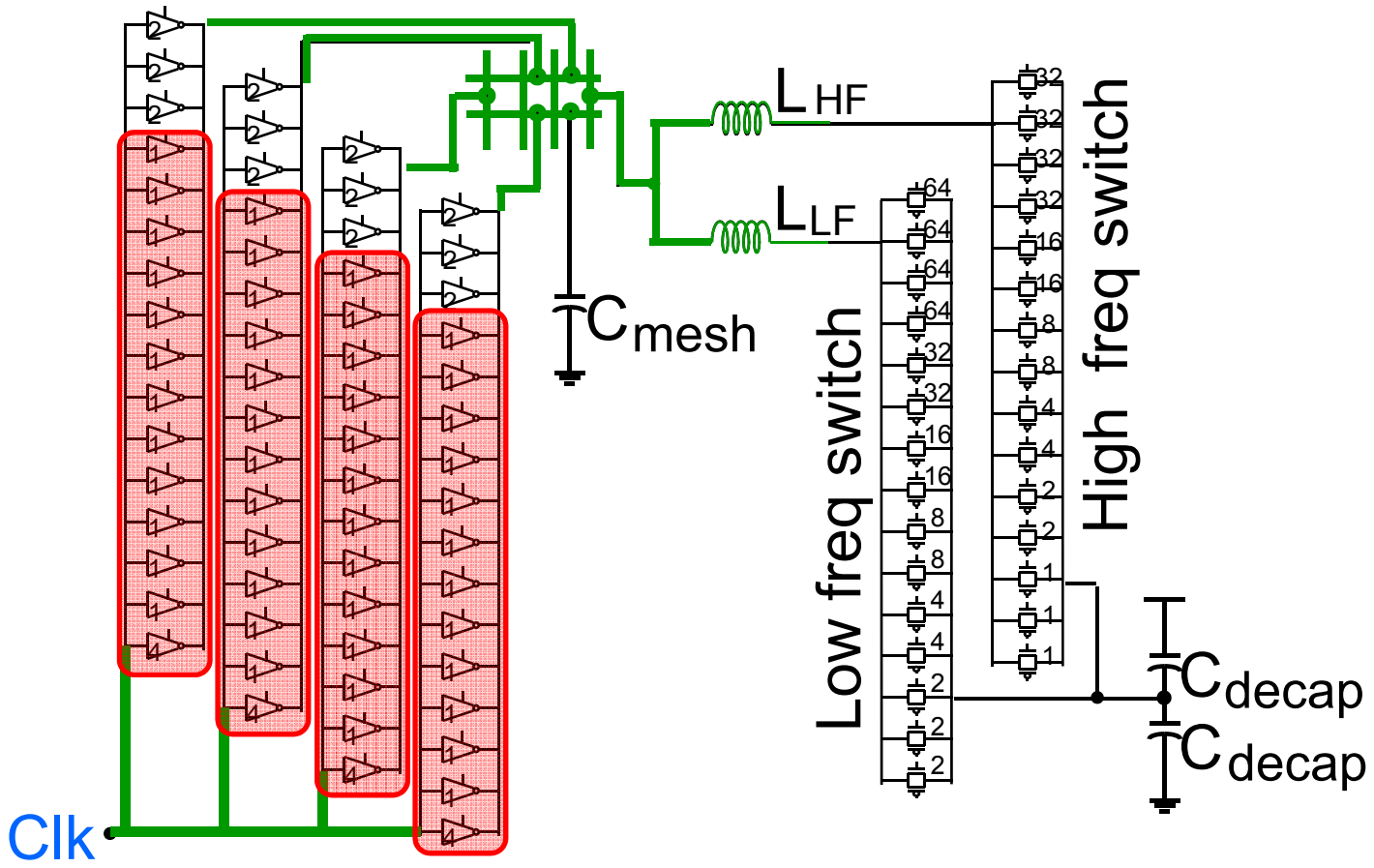
On-The-Fly Mode Changing

The Sector Buffers and Switches have 16 strength setting to enable On-The-Fly mode changes



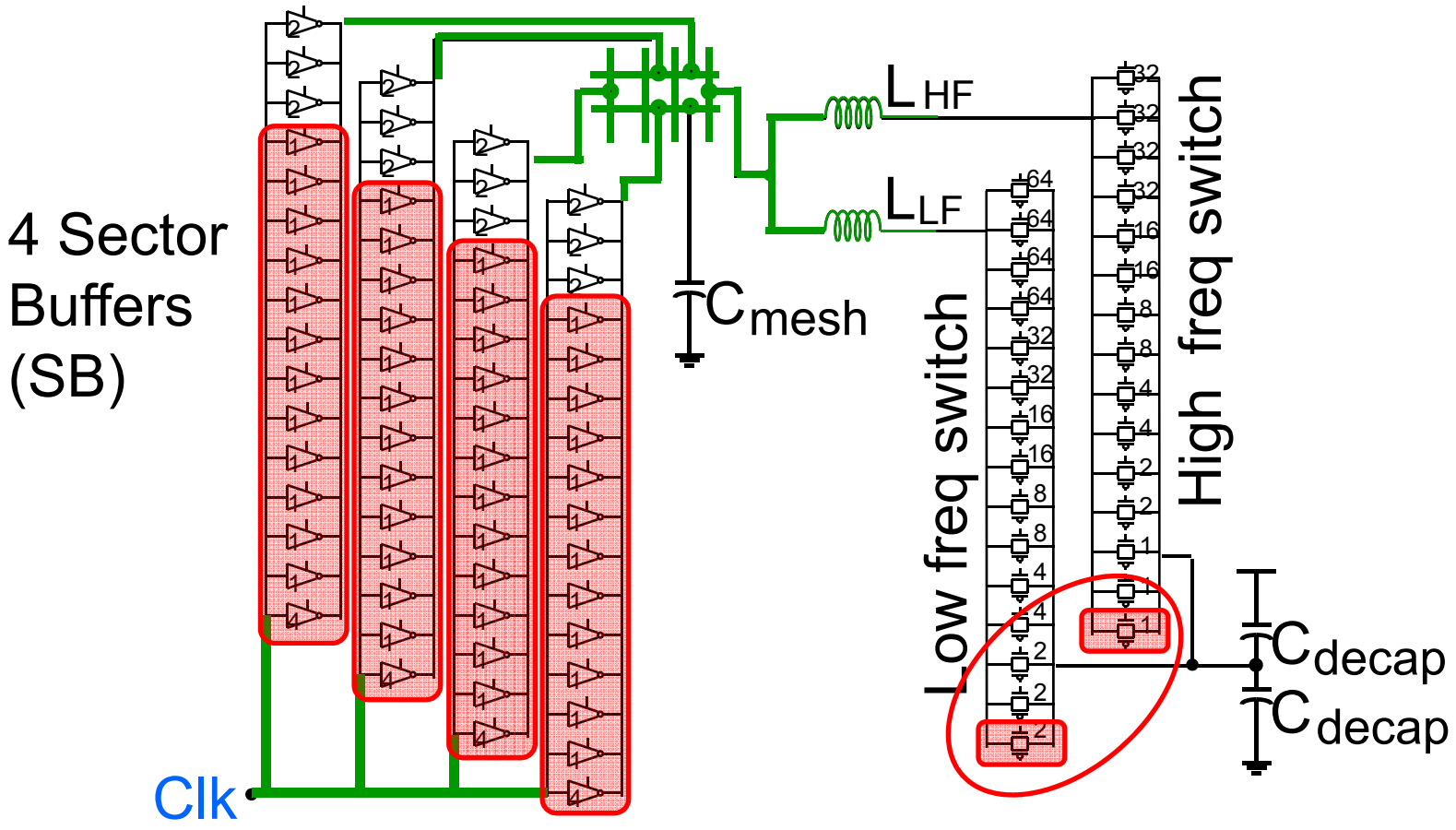
Circuits in One Sector

Highlighting devices **on** for NR_{clk} (Non-Resonant Mode)



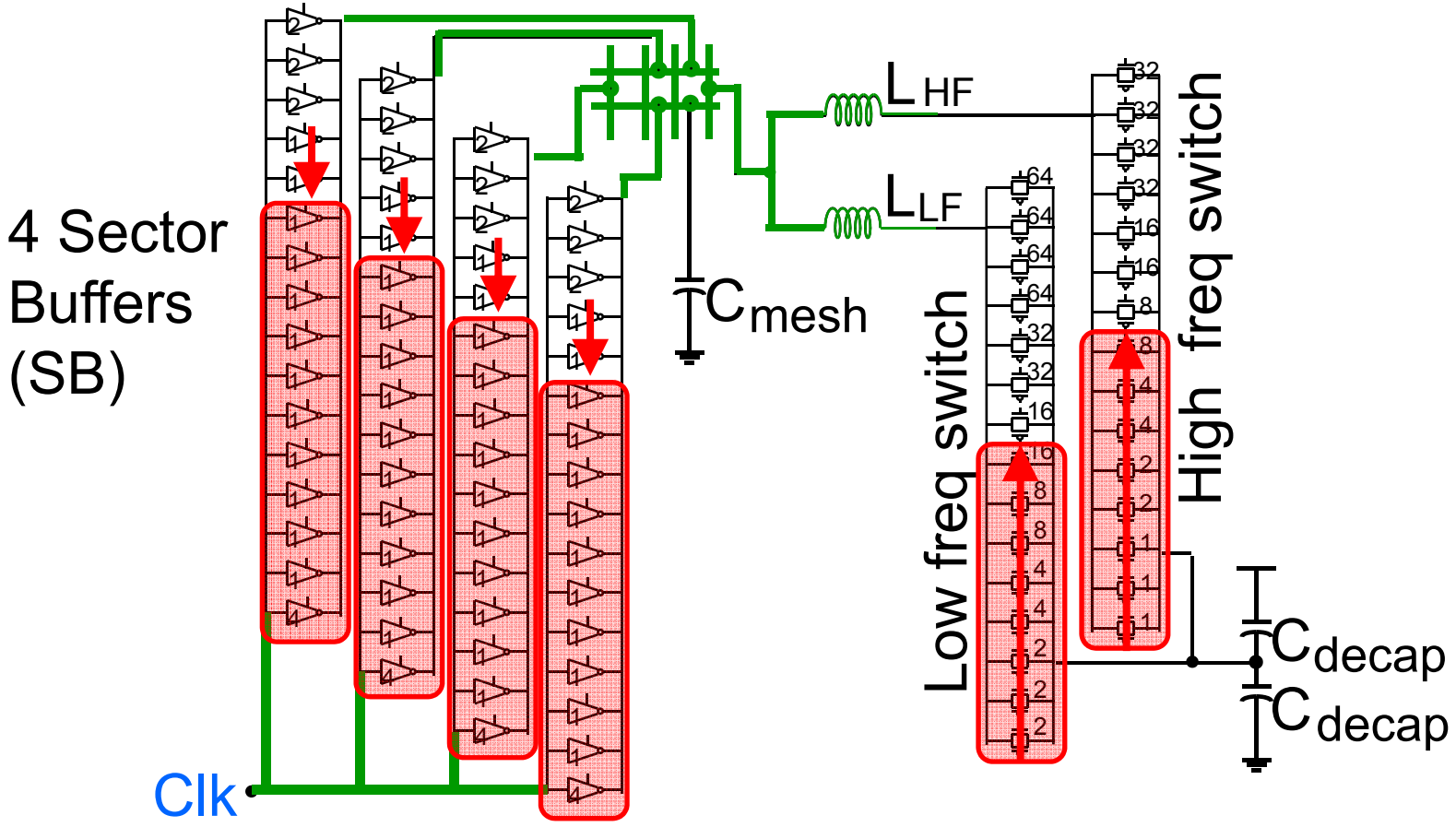
First Mode Changing Step

- Take first step toward HF_{clk} by closing smallest part of each switch



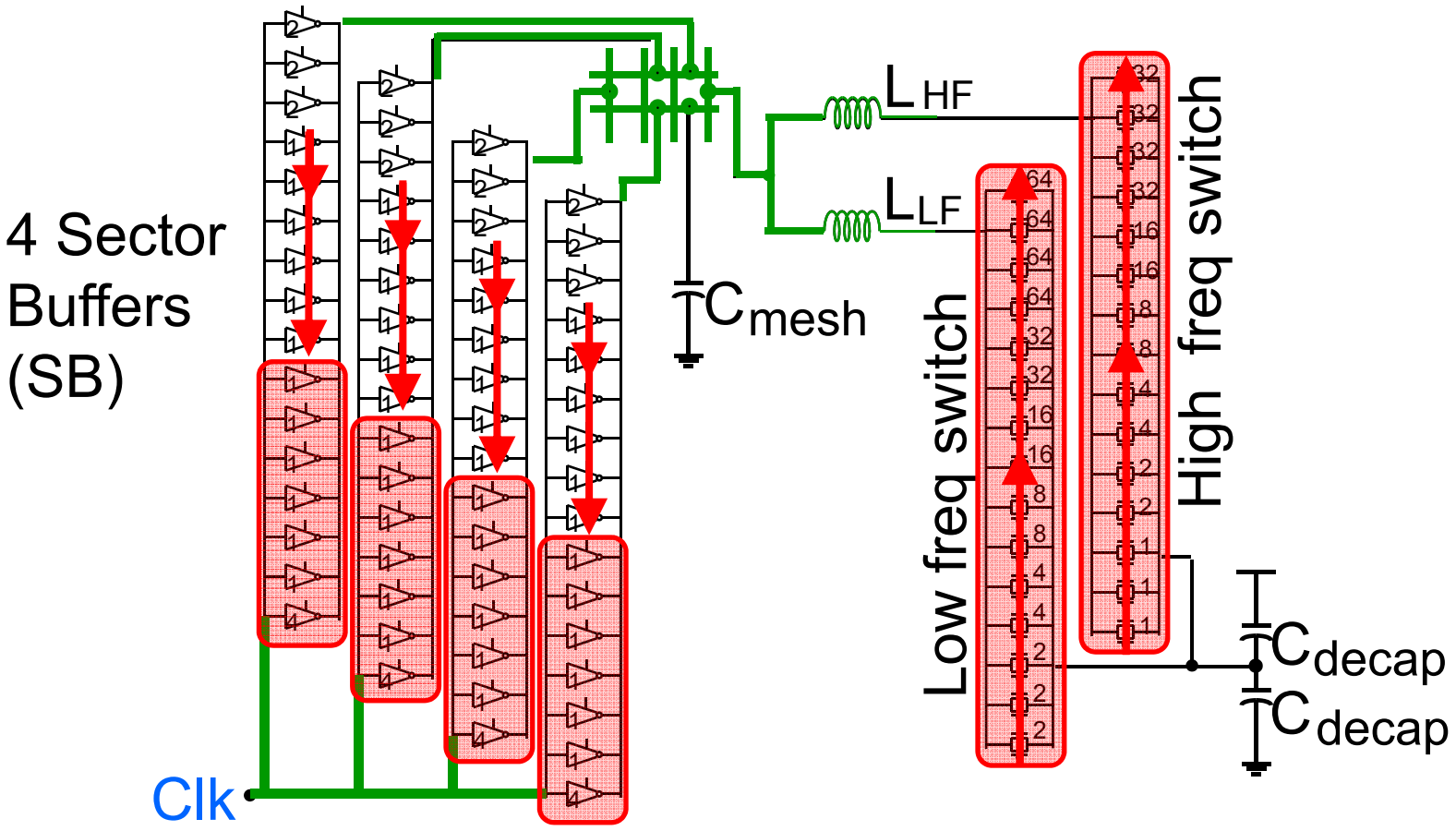
Mode Changing

- Halfway from NR_{clk} to HF_{clk}
  - closed more switches, and reduce sector buffer drive strength



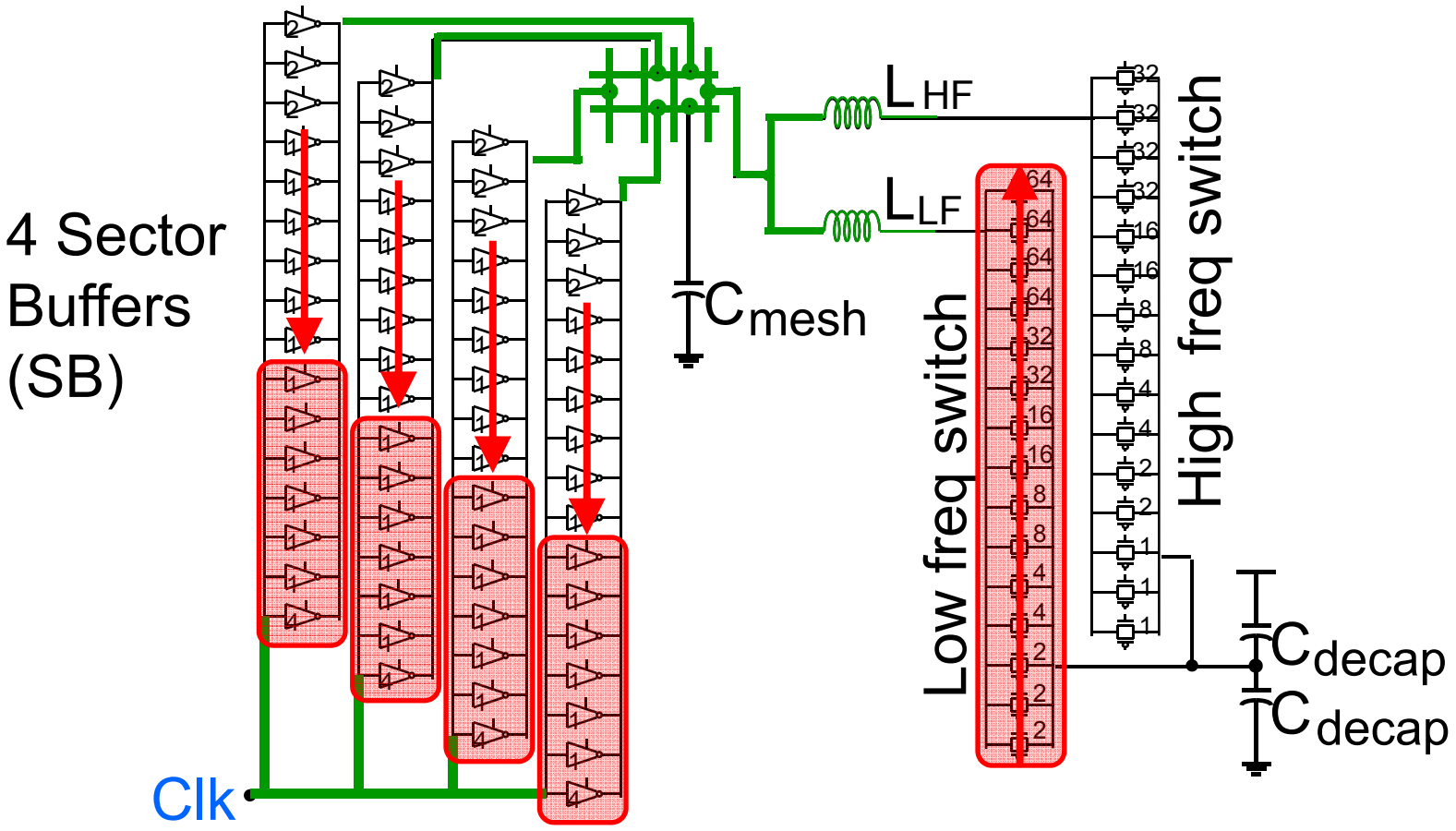
Completed Mode Changing

Highlighting devices **on** for HF_{clk}



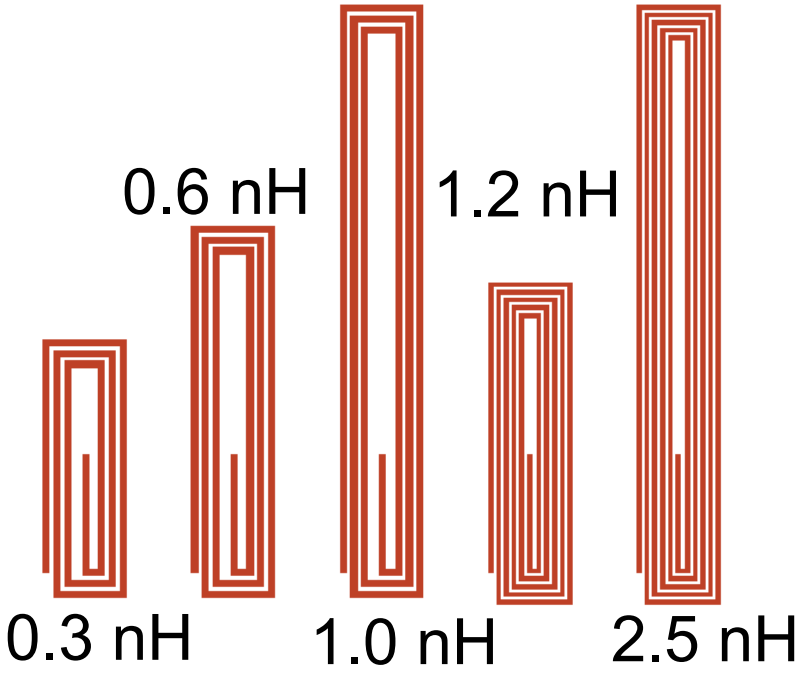
Low Frequency Resonant Mode

Highlighting devices **on** for LF_{clk}

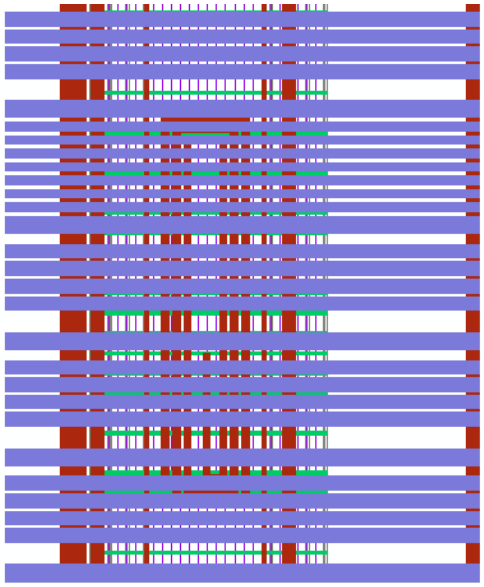


Inductors

- 13 Inductors from 3 to 5 turns and 0.3 to 2.5 nH available
  - Inductors use only one level of UTM (2nd from top)
  - Width constrained to minimize UTM blockage
  - Can be placed over any digital circuit

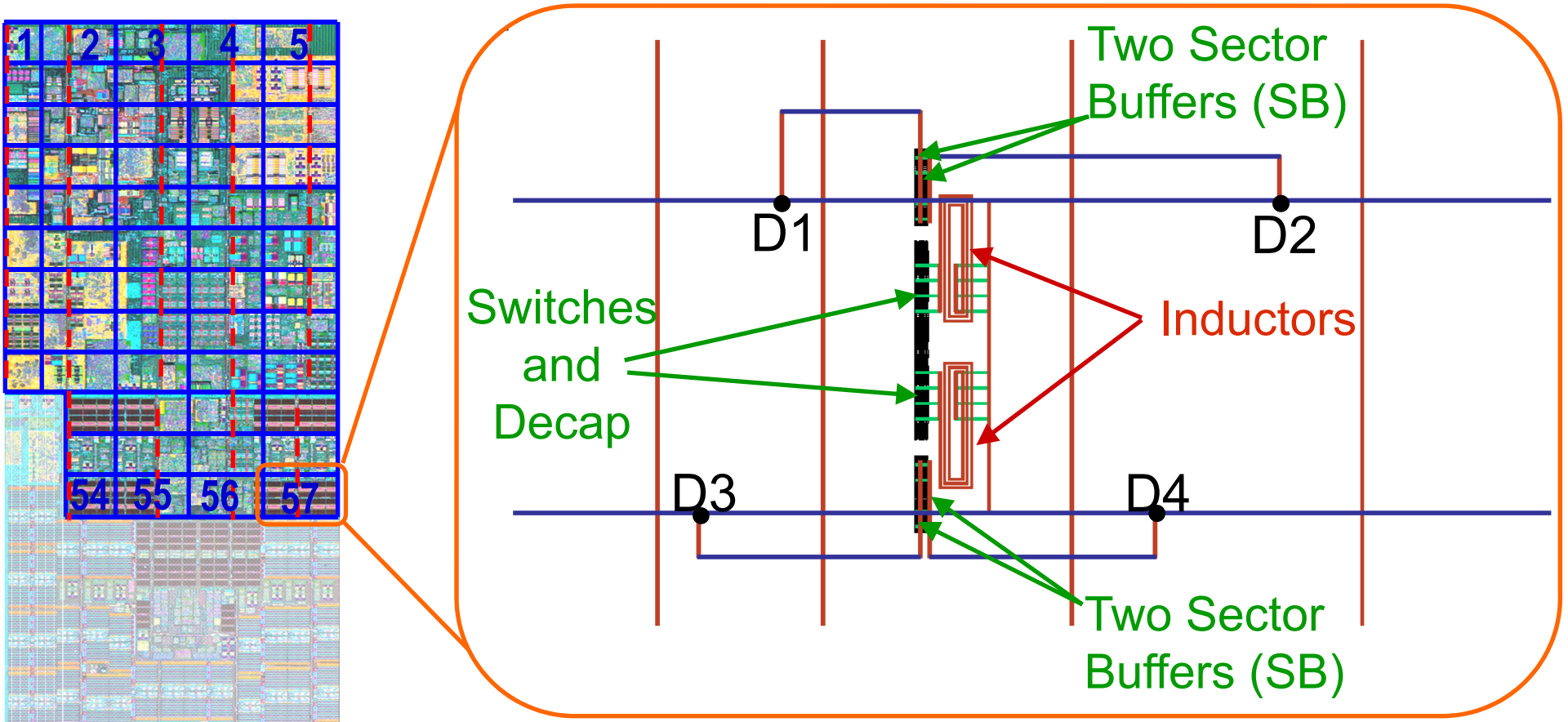


0.6 nH inductor in power-mesh environment

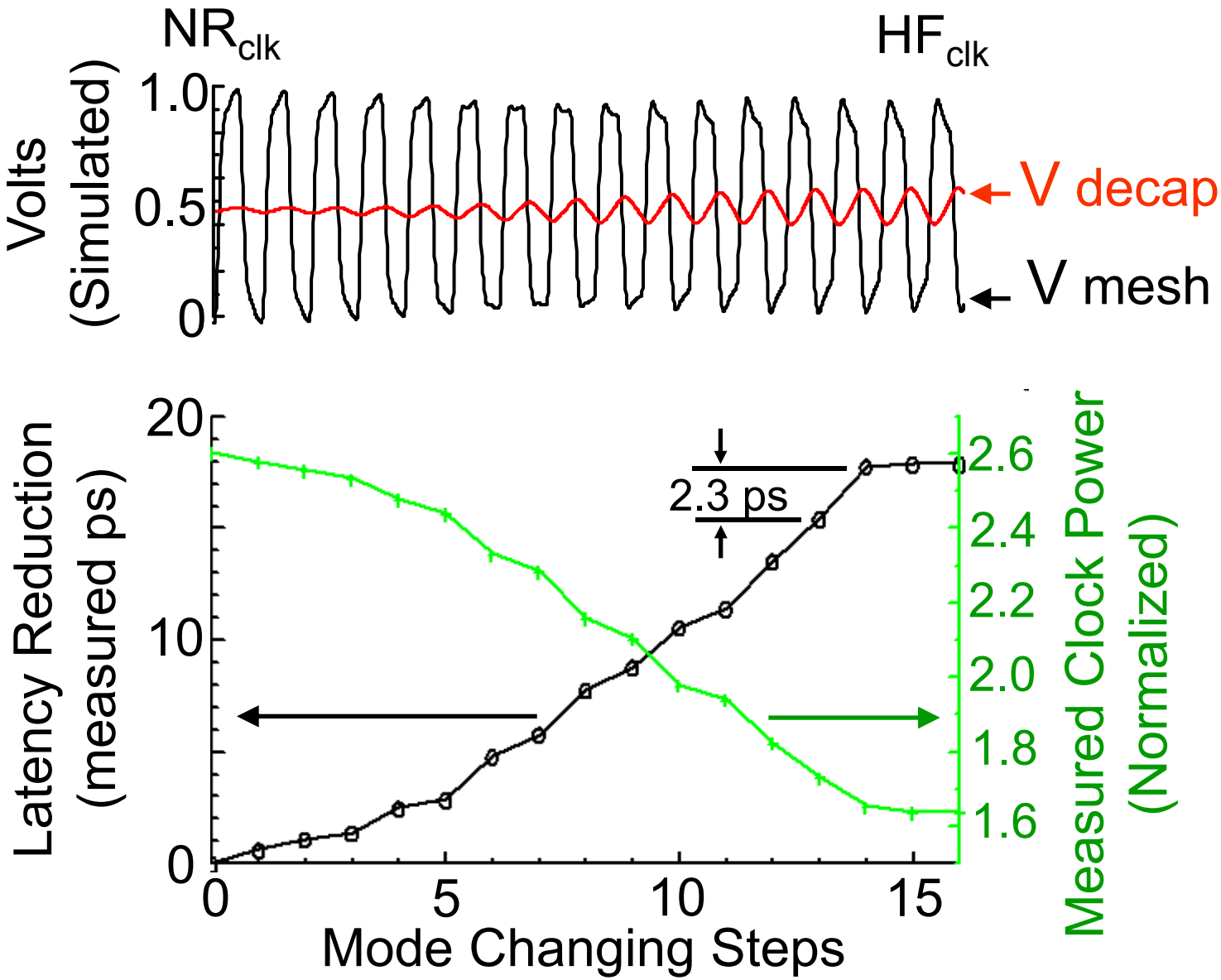


One Resonant Sector Layout

- Inductors are placed independent of Buffers, Switches, and Mesh wires
- Sector buffer drive points (D1 to D4) chosen based on loads

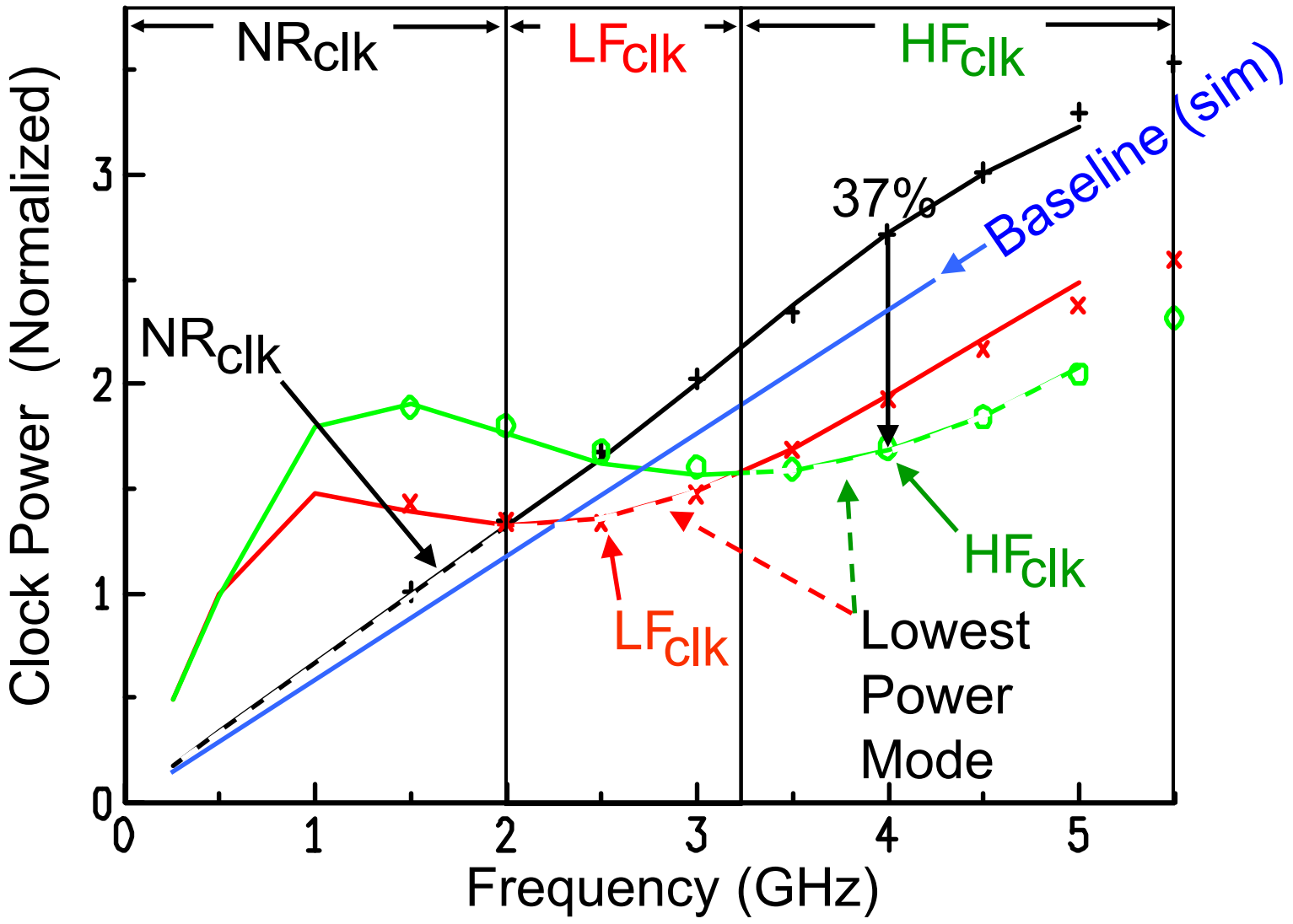


Gentle Mode Changing Results



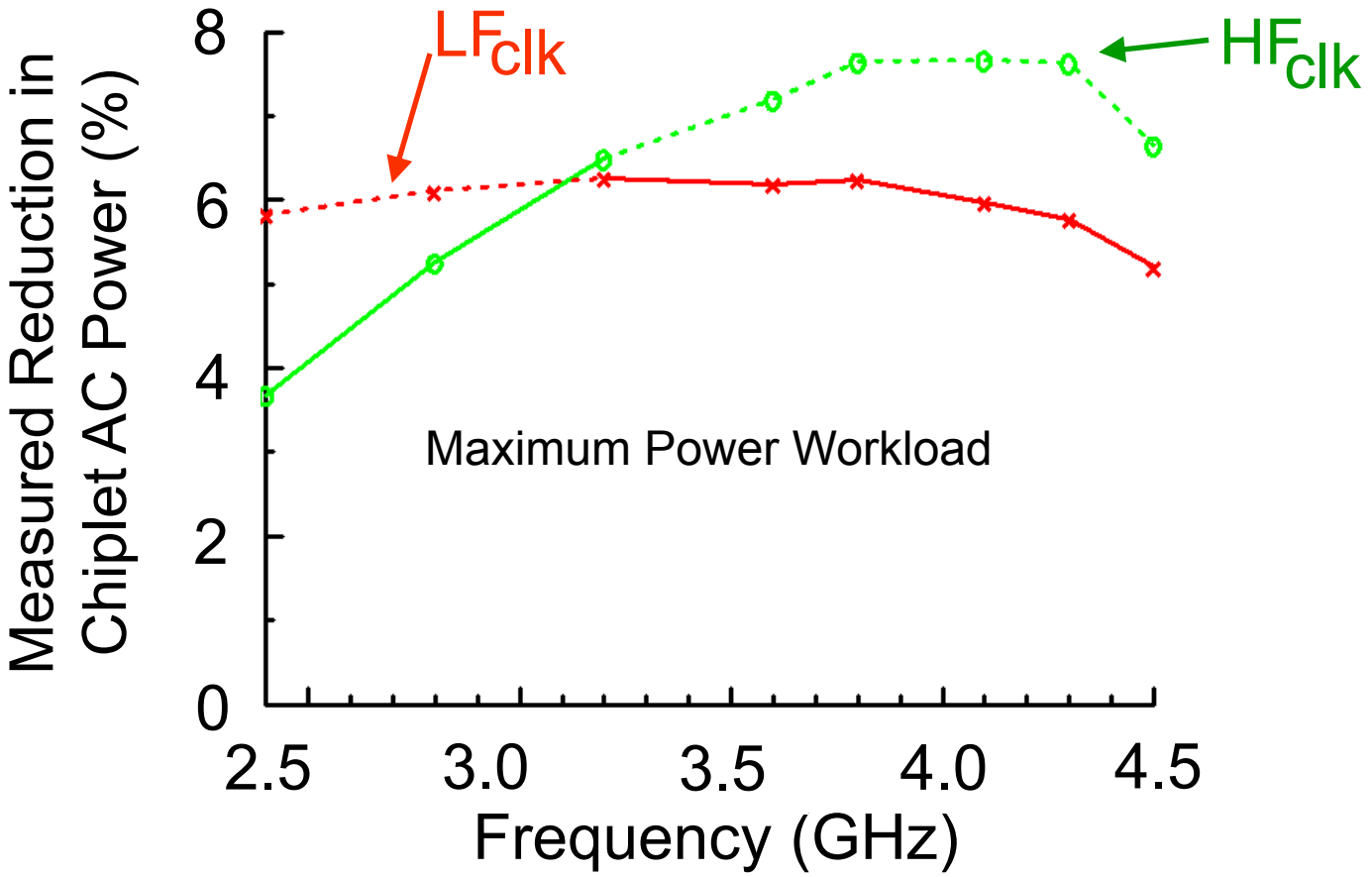
Power vs. Frequency

- Measured and simulated power at constant VDD



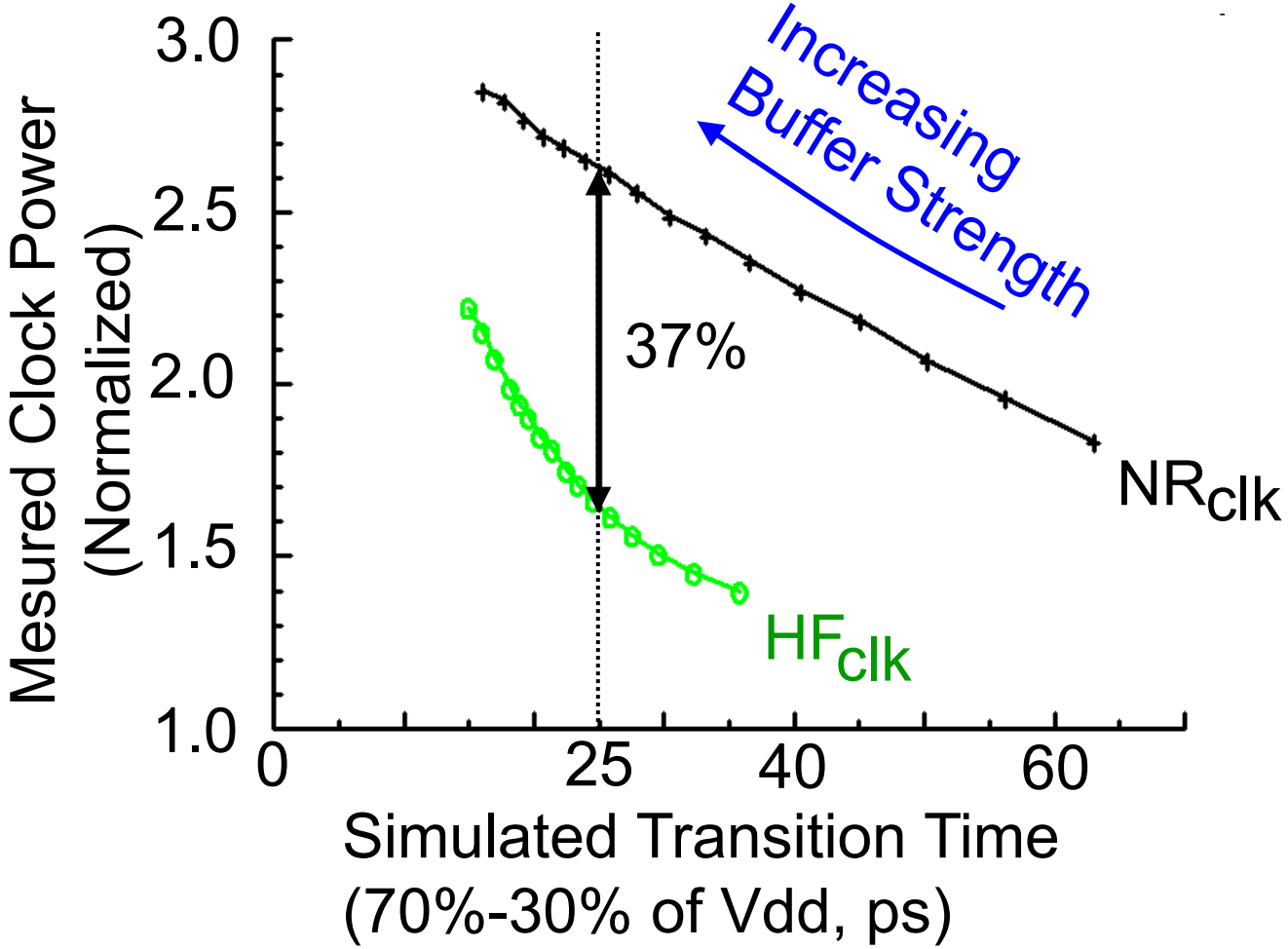
Resonant Power Reduction

- Voltage scaled with Frequency
  - Using integrated voltage regulators
- Corresponds to ≥ 4% of total chip power



Clock Signal Quality
Power vs. Transition Time

- Buffer strengths chosen to give 25ps transition time



Summary

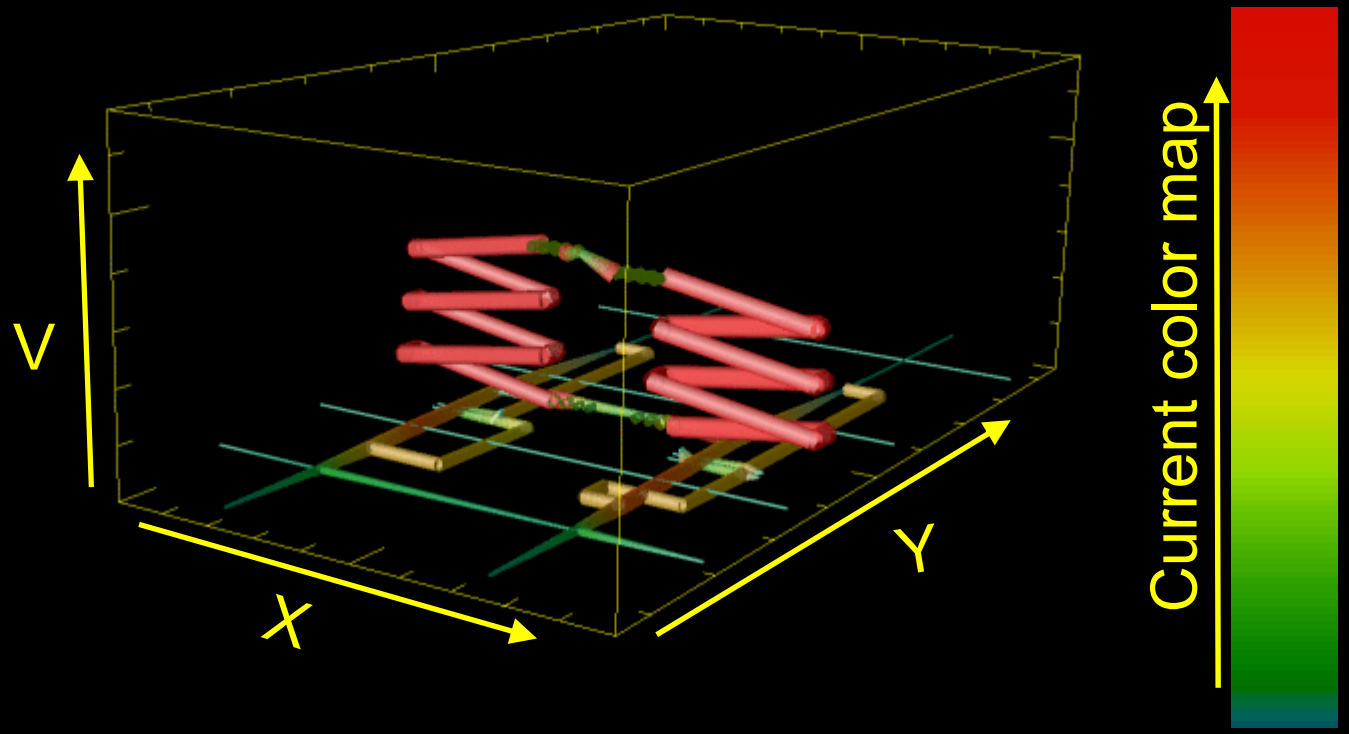
- A wide-frequency-range resonant clock was implemented saving power from 2.5 GHz to more than 5 GHz
- On-the-fly mode changing with an on-chip controller keeps each chiplet in the lowest power mode as the voltage and freq. of each chiplet is optimized
  - We measured a global clock resonant power reduction of 37% corresponding to 28% savings compared to a non-resonant baseline, or 33% savings of the sector buffer power used to drive the clock meshes
  - This measured power reduction is roughly equivalent to the AC power of one of the 12 cores

Acknowledgments

- We would like to thank
  - The many IBMers across the world from technology through design, test, and bring-up, who collaborated to make this work successful
  - Our leadership for support from research to product

Voltage & Current Visualization

One sector in HF_{clk} mode showing wires and circuits



Diameter & Color shows currents

Ivytown: A 22nm 15-core Enterprise Xeon® Processor Family

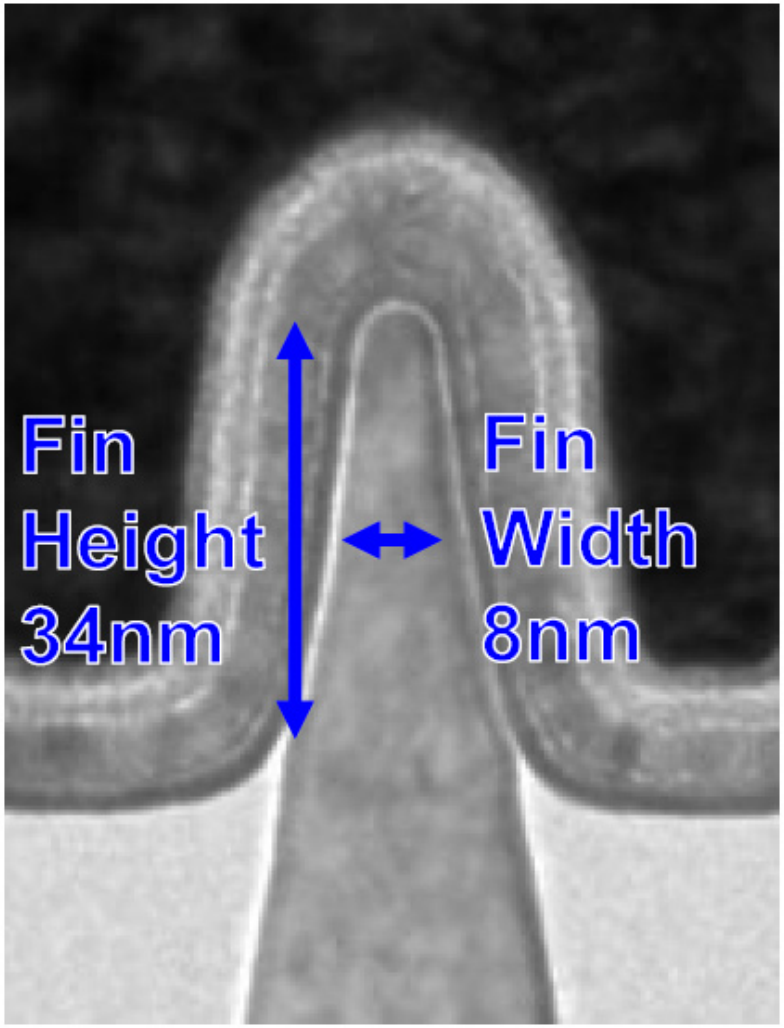
Stefan Rusu, Harry Muljono, David Ayers, Simon Tam, Wei Chen, Aaron Martin, Shenggao Li, Sujal Vora, Raj Varada, Eddie Wang

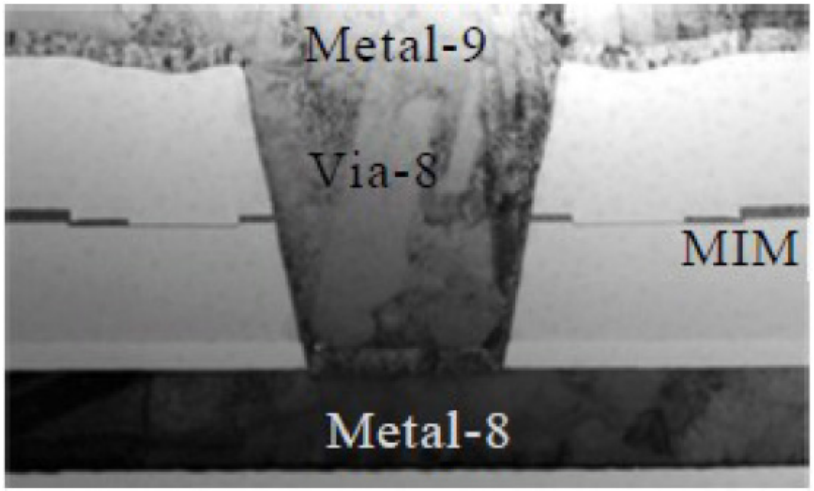
Intel Corporation, Santa Clara, CA

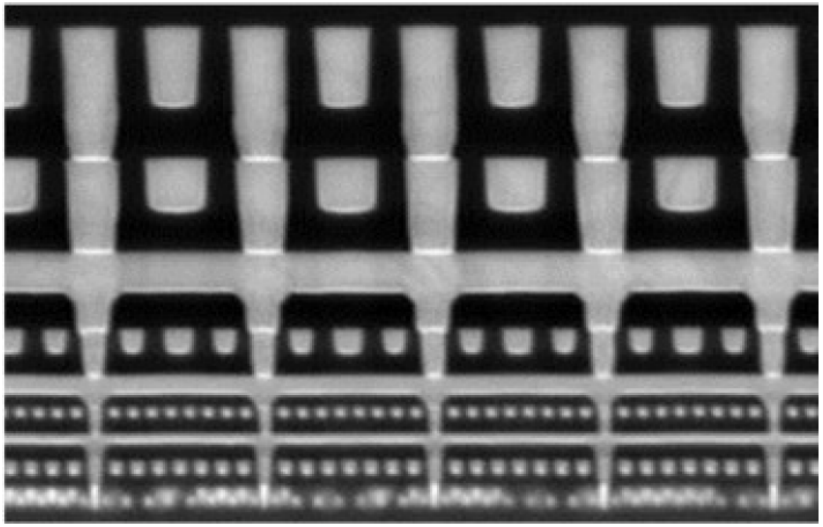
Outline

- Process Technology
- Block Diagram
- Floorplan and Die Photo
- Cache Design
- Clock and Voltage Domains
- Core, Cache and IO Recovery
- Power, Package and Thermals
- DDR / VMSE and Serial Links
- DFT and DFM Features
- Summary

Process Technology





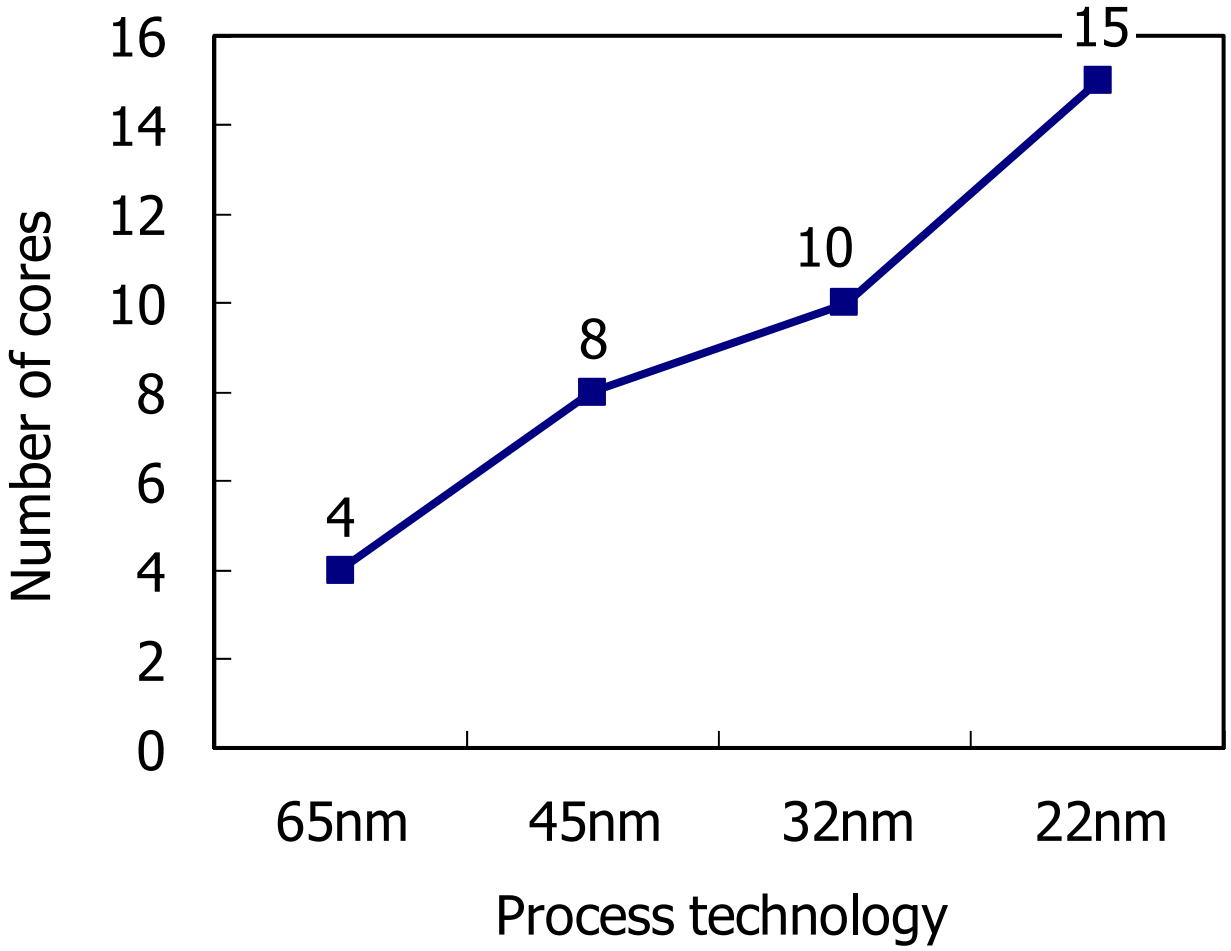


- 22nm Tri-gate transistors

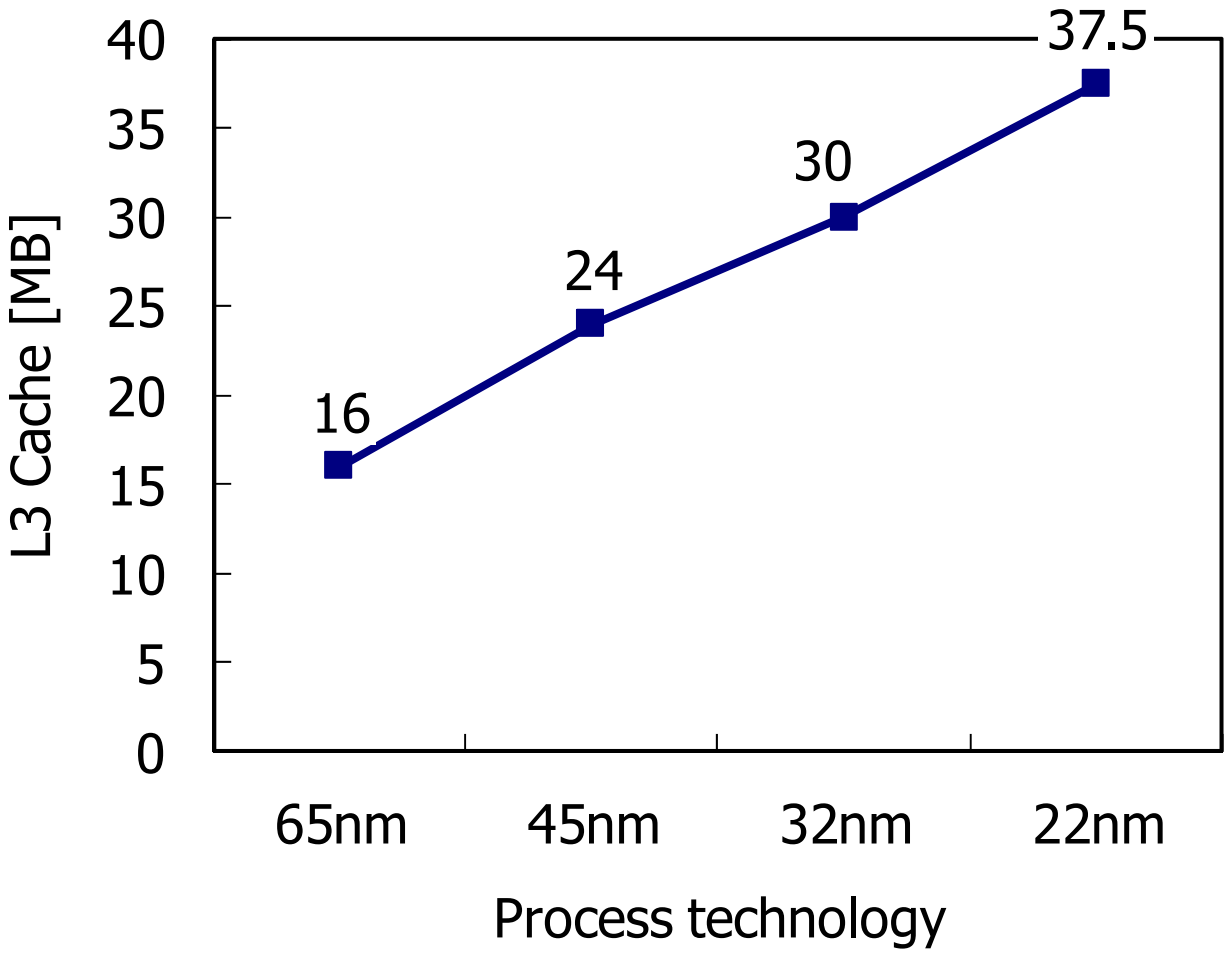
[C. Auth, et al., VLSI Symposium 2012]

- 9 metal layers with integrated MIM capacitor

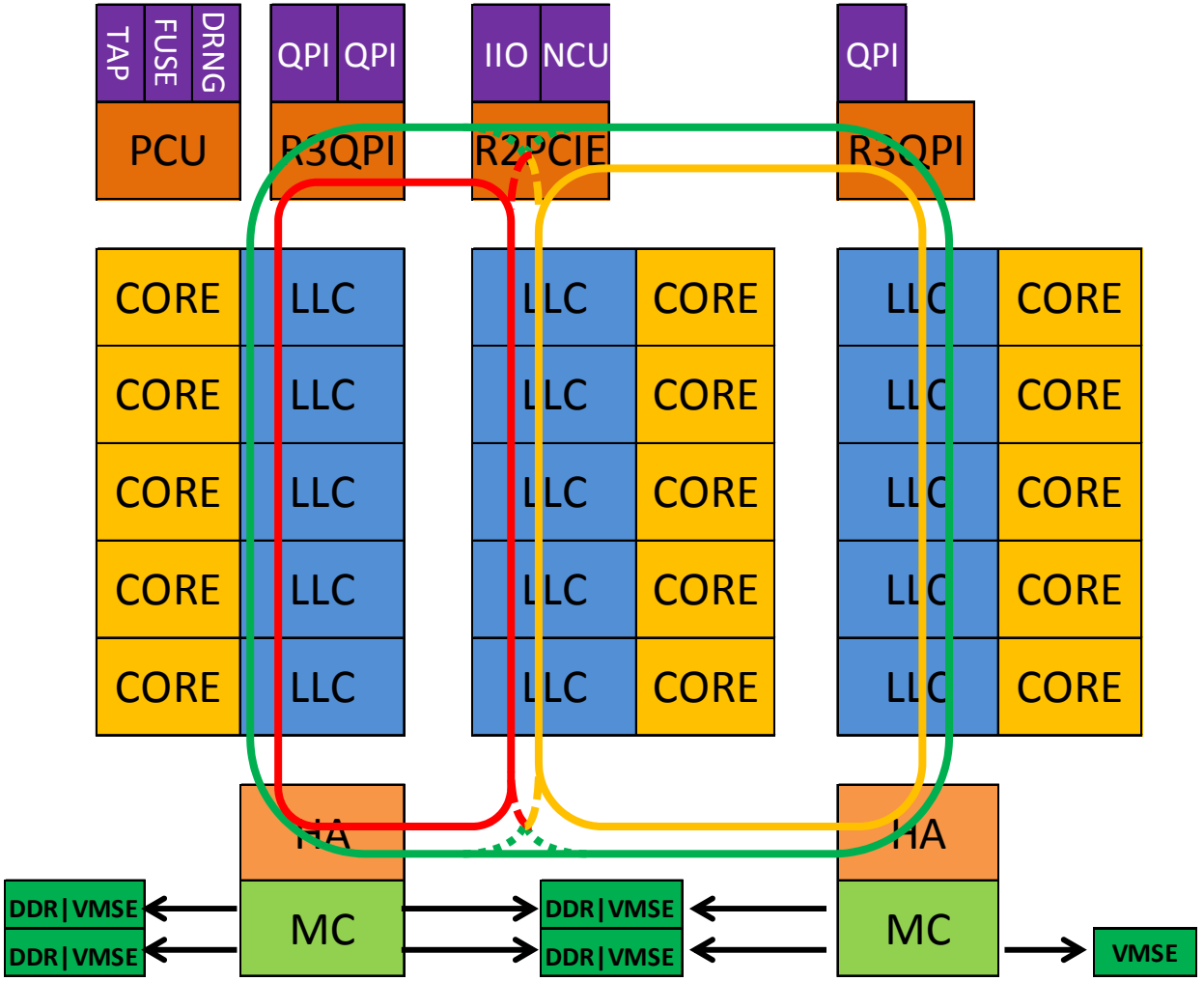
Xeon® Processors Core Scaling



Xeon® Processors L3 Cache Scaling

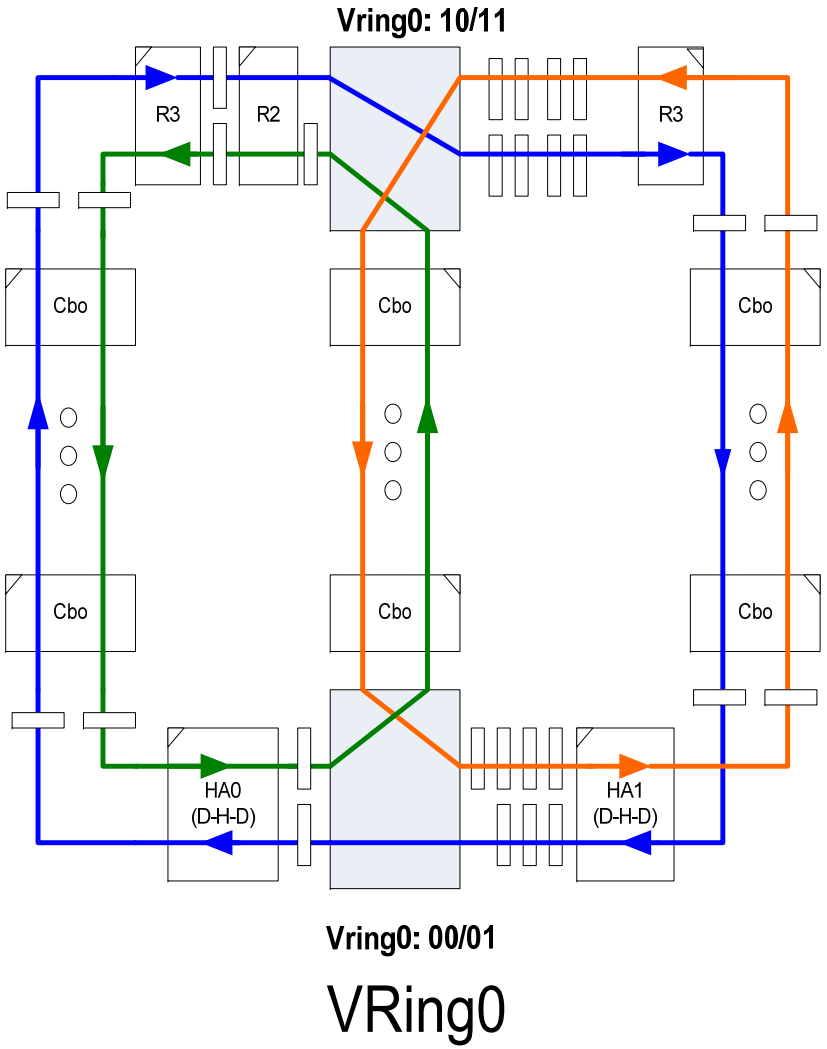


Processor Block Diagram

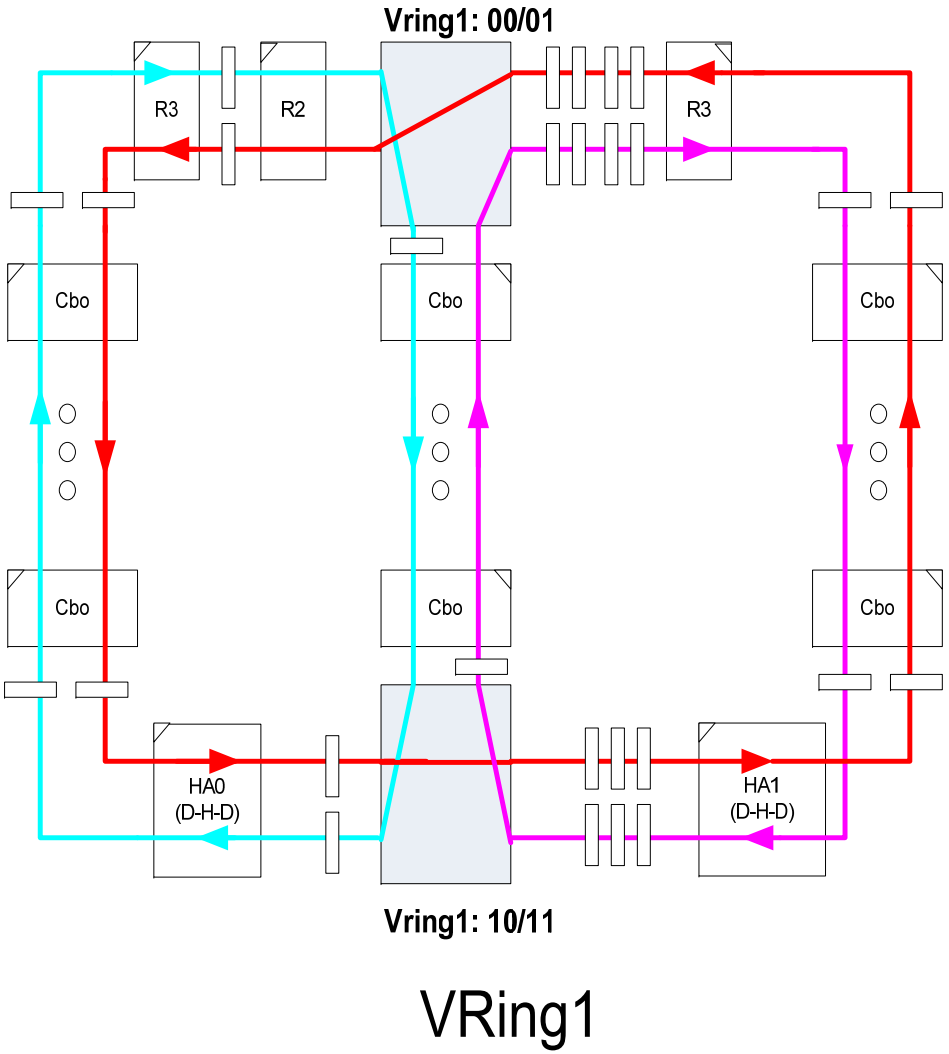


- 15 cores, 30 threads, 2 integrated memory controllers

Ring Operation

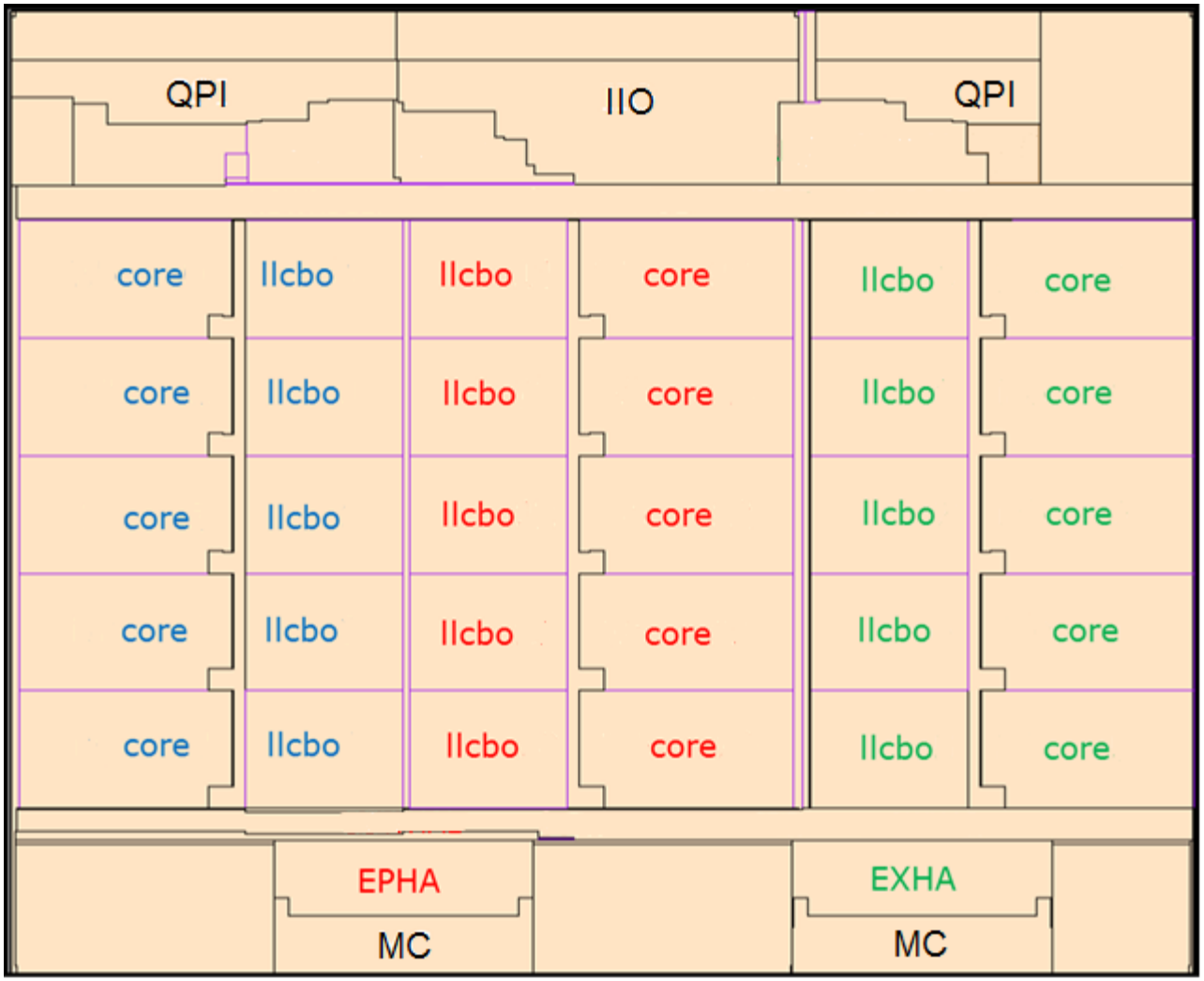


VRing0

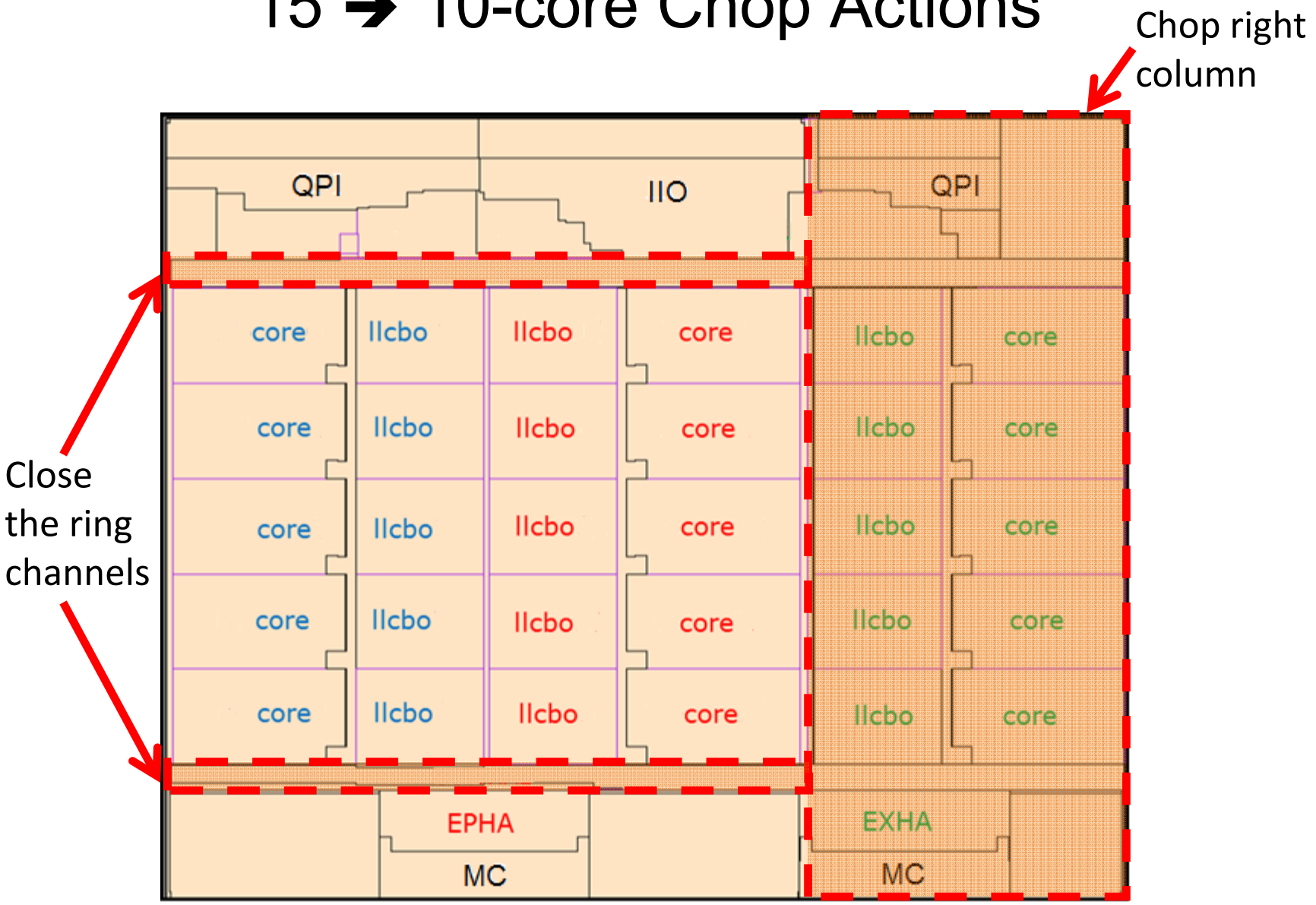


VRing1

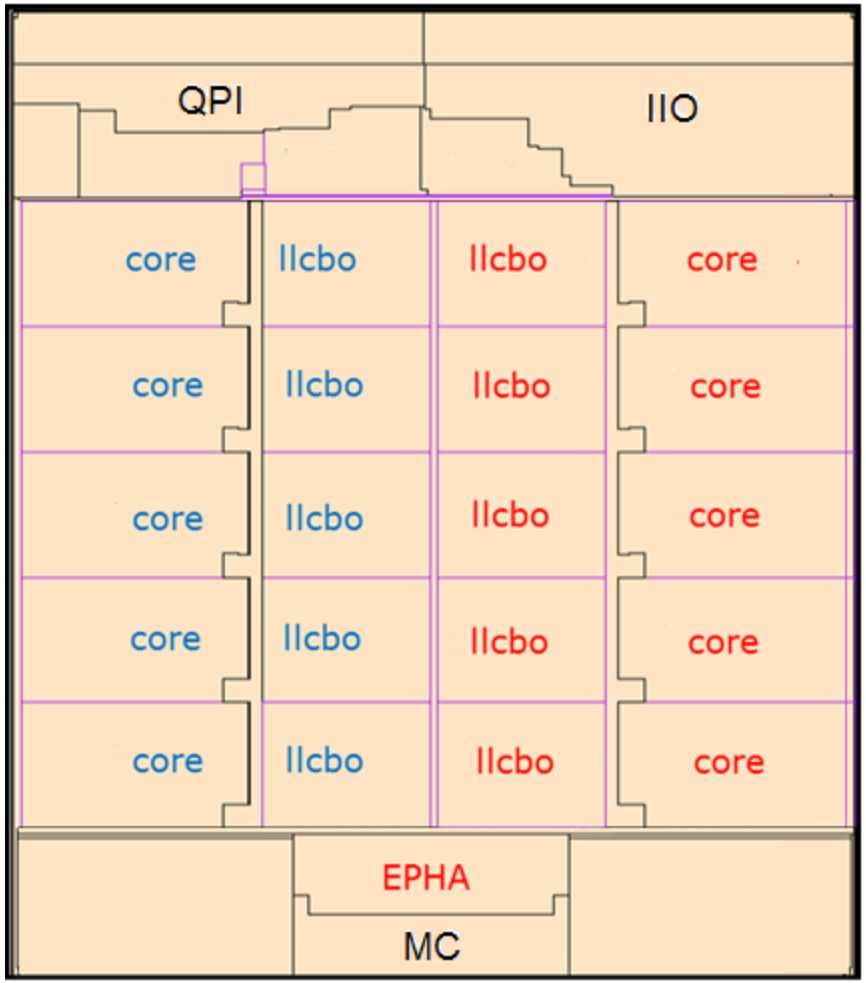
15-core Floorplan



15 ➔ 10-core Chop Actions

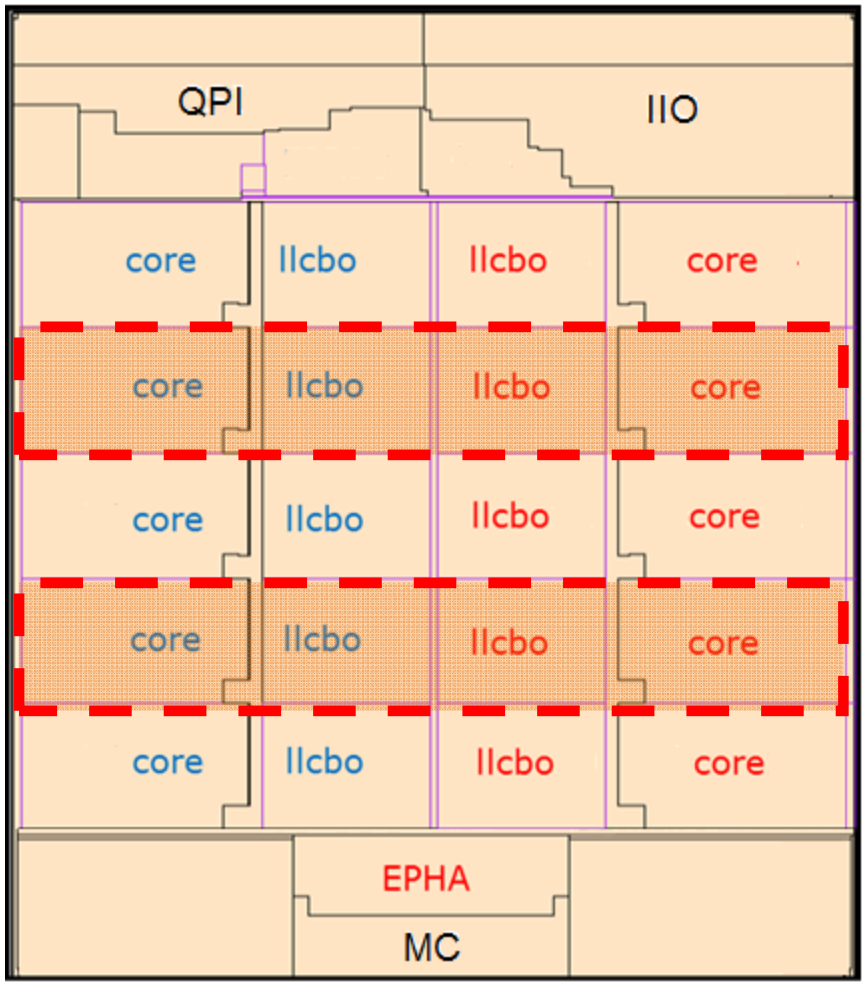


10-core Floorplan

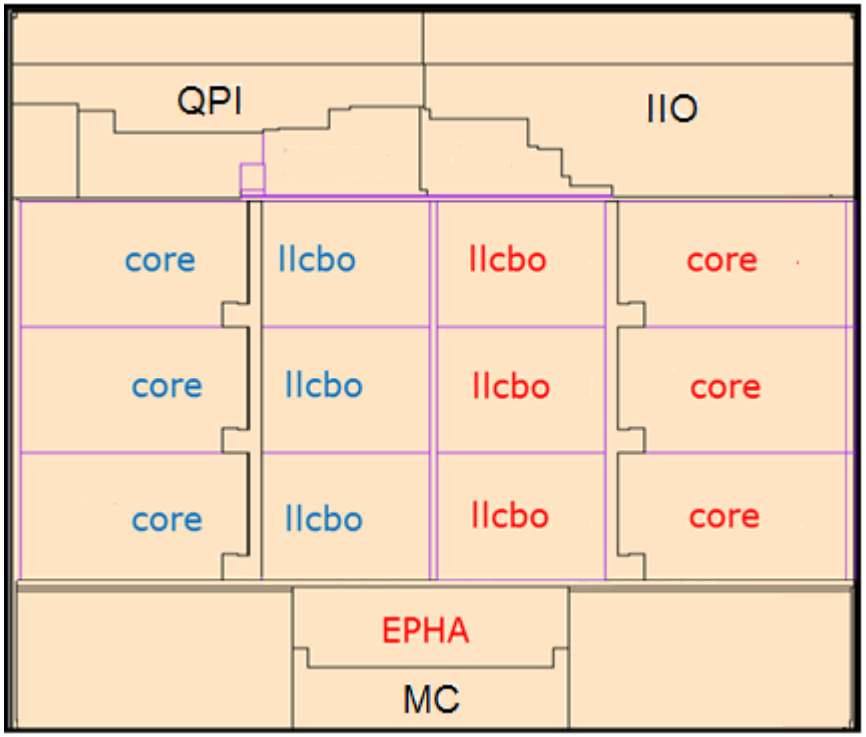


10 ➔ 6-core Chop Actions

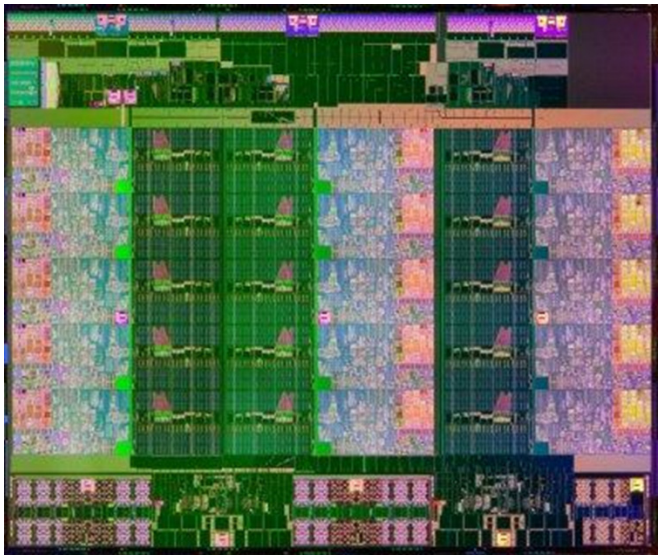
Remove
2nd and
4th rows

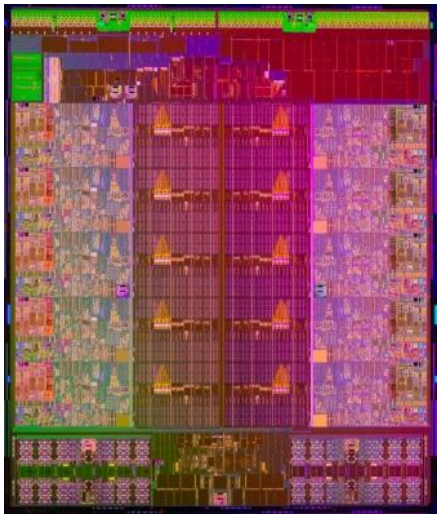


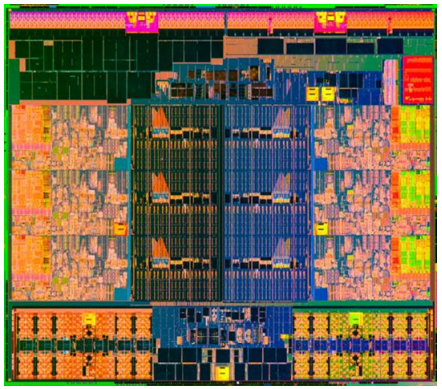
6-core Floorplan



Ivytown Processor Family

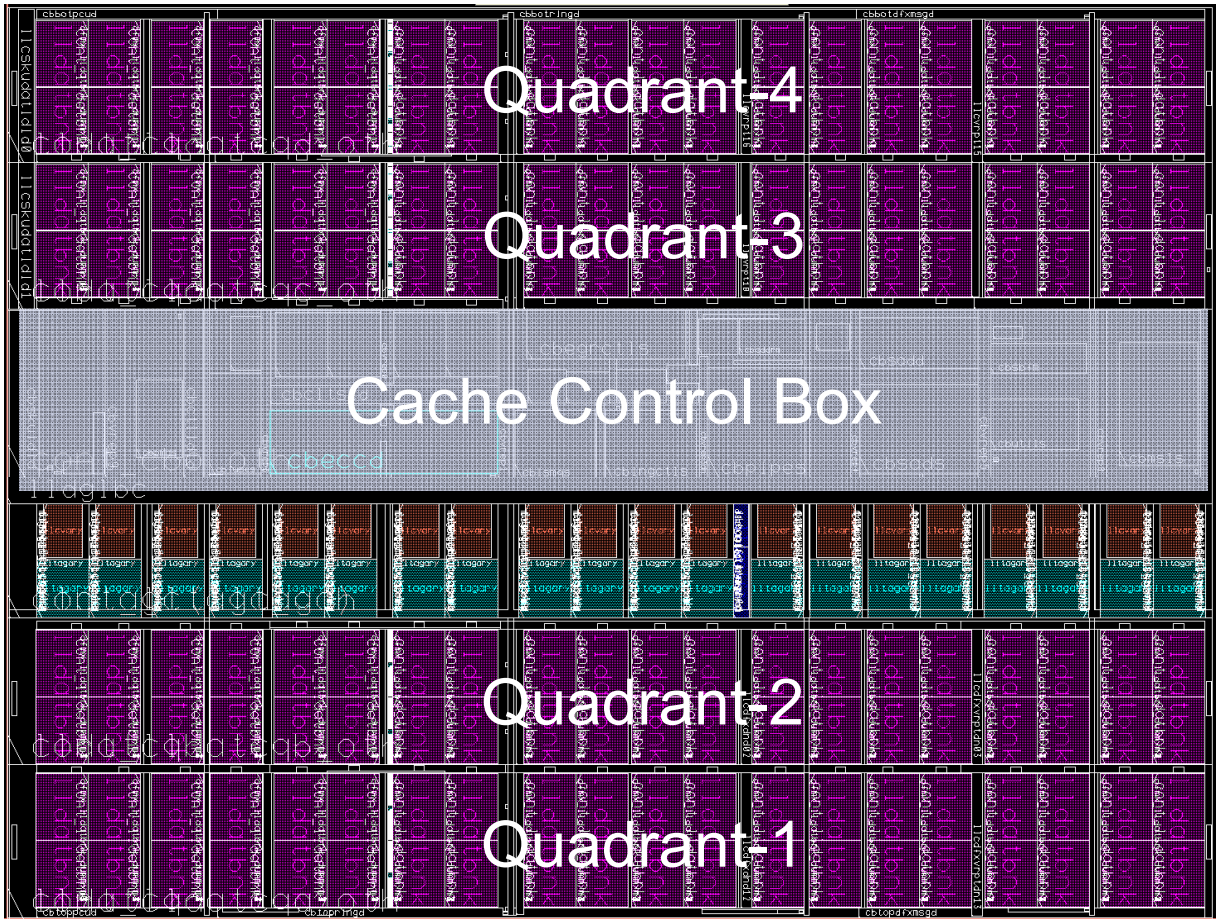






Cores:	15	10	6
L3 cache [MB]:	37.5	25	15
Transistors [B]:	4.31	2.89	1.86
Die size [mm^2]:	541	341	257

L3 Cache Slice



Bit cell area [μm^2]

Data arrays ► 0.108

CVLRU arrays ► 0.170

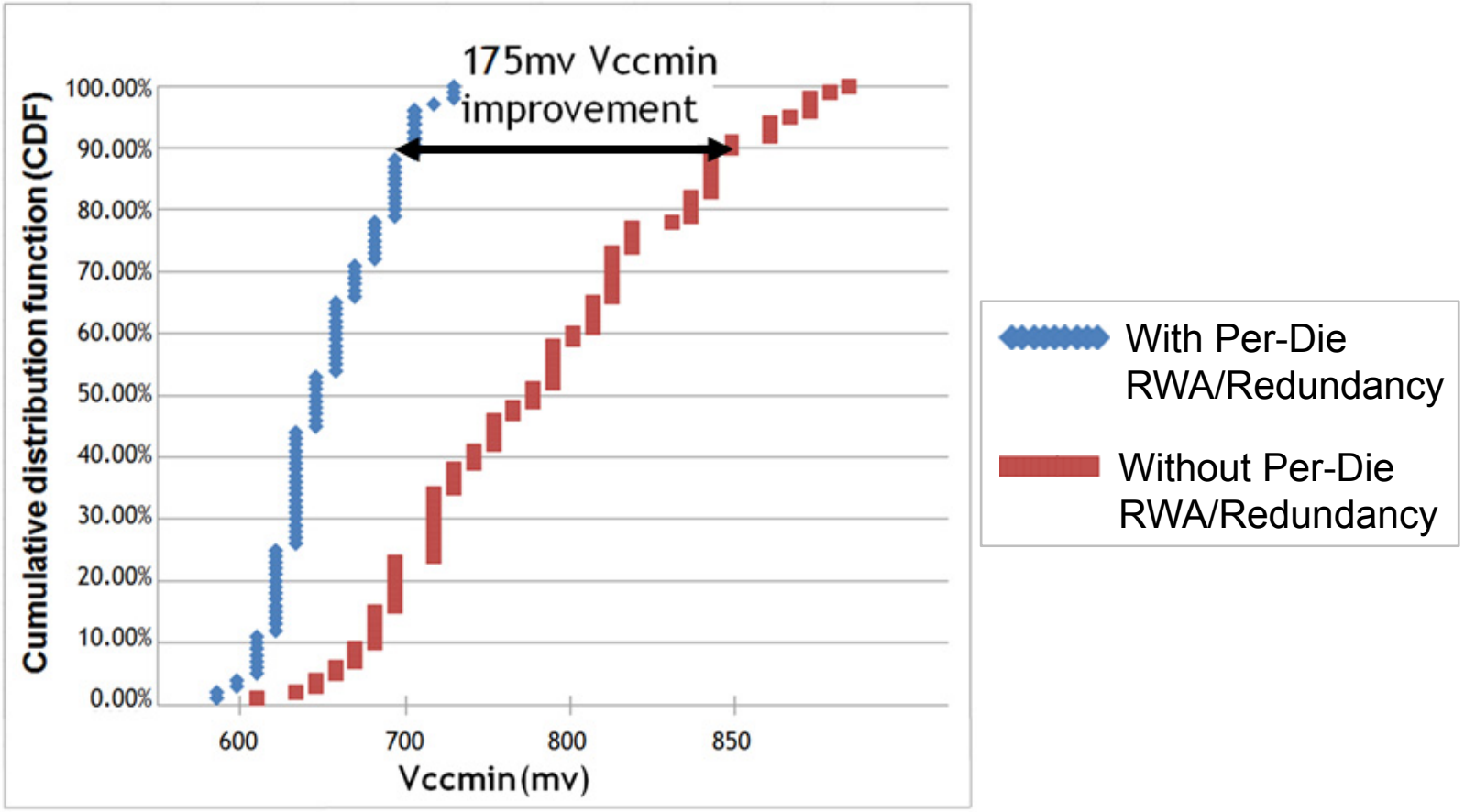
Tag arrays ► 0.130

Data arrays ► 0.108

80 data arrays, 2048 sets and 20 ways

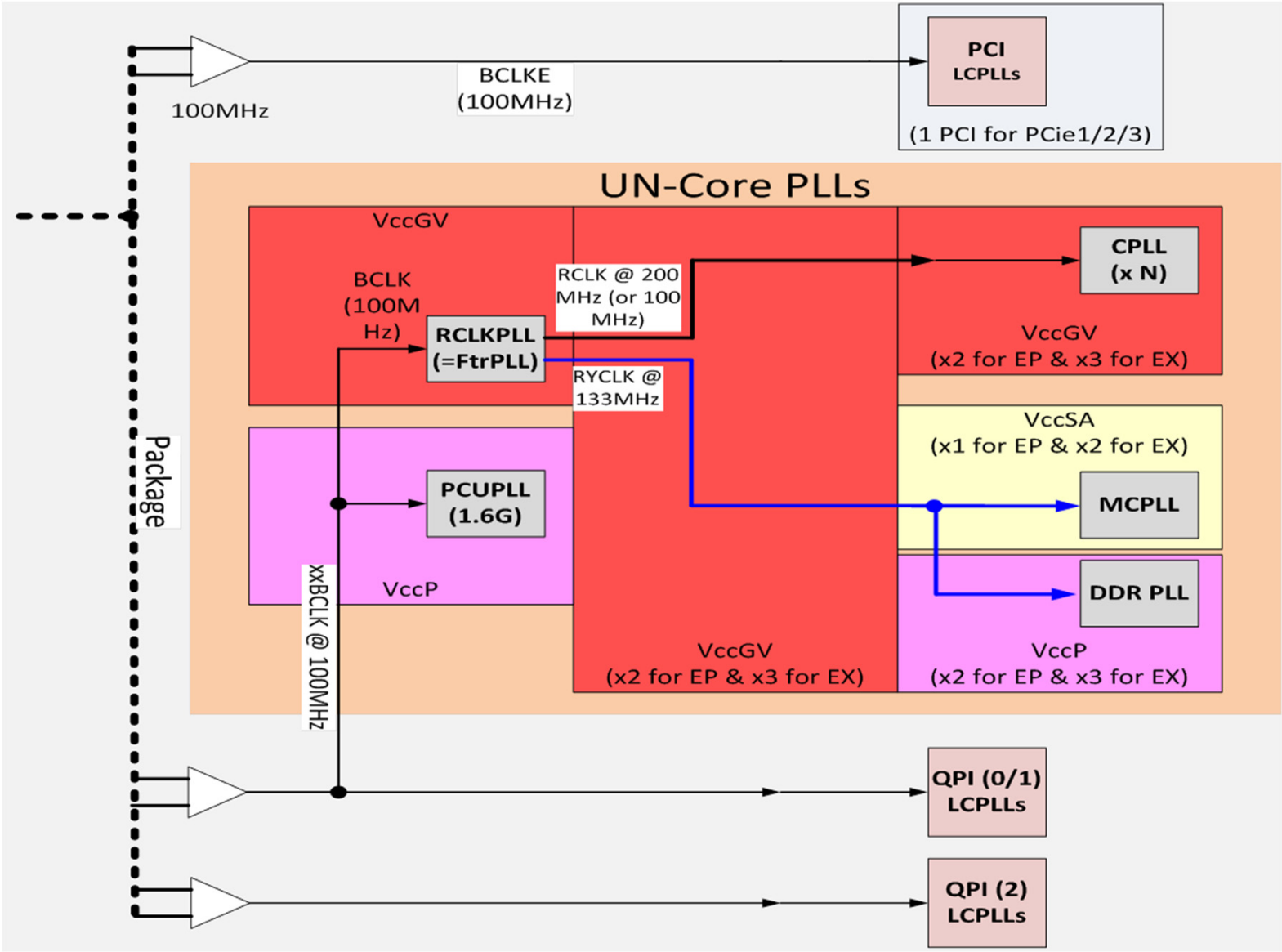
DECTED in data arrays, SECDED in tag and CVLRU arrays

L3 Cache 40MB Vmin Improvement at 1.2GHz

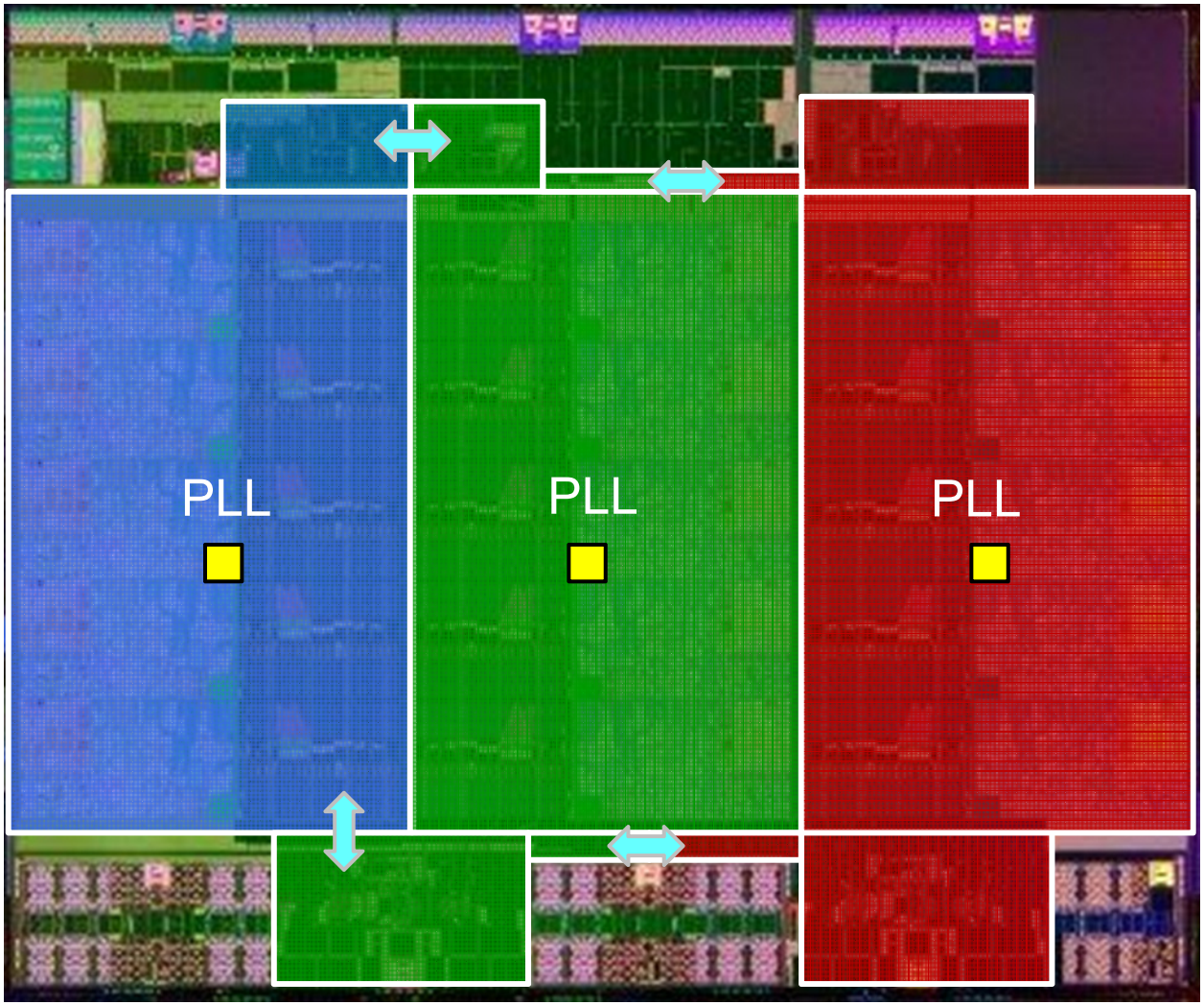


- Achieved 680mV Vmin at 90% tile
- 175mV Vccmin improvement from per-die redundancy, per-die RWA and DECTED

Clock Distribution

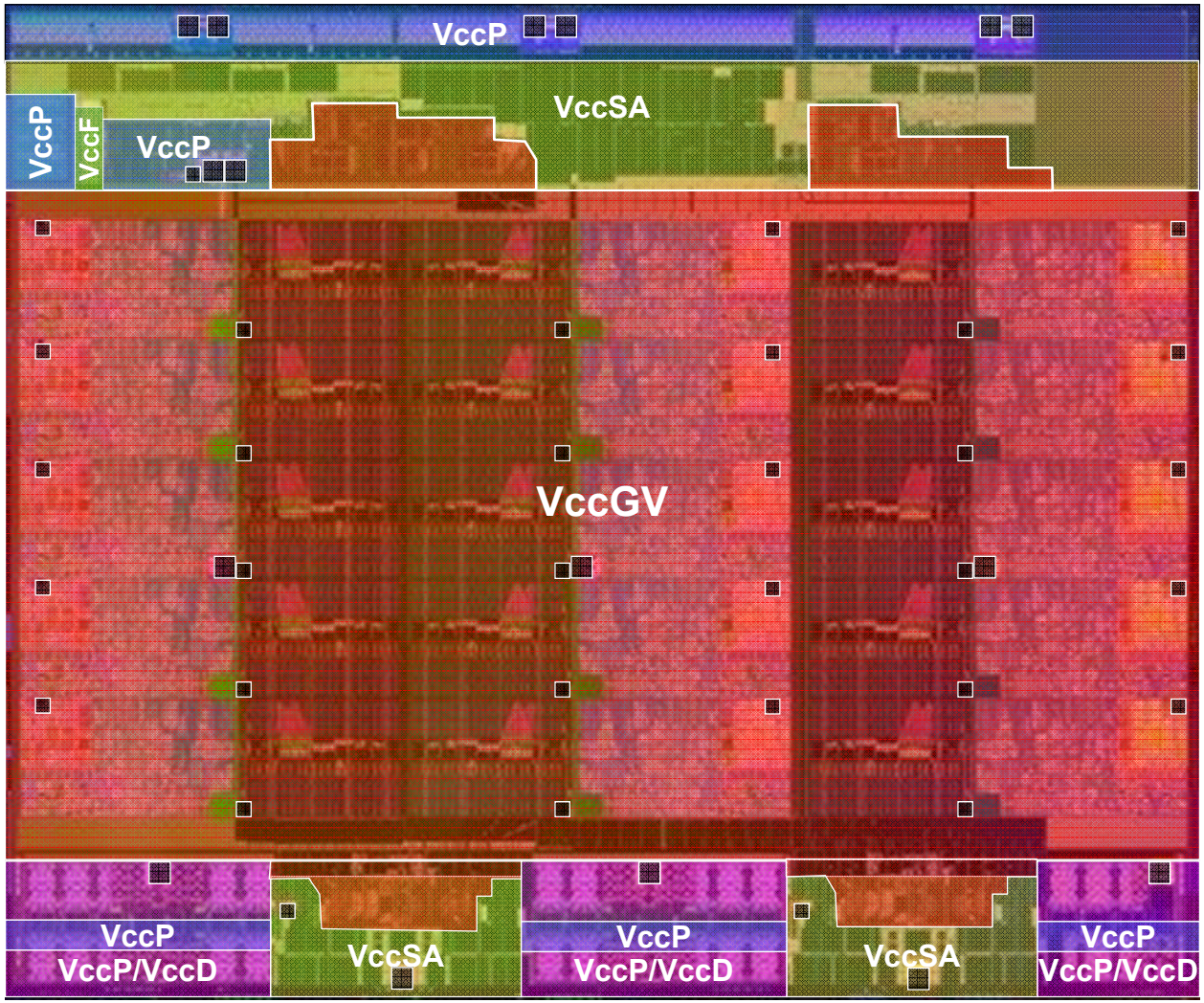


Clock Domains



Single PLL per column reduces power and skew

Voltage Domains



VccGV supplies cores, cache, and ring (red)

- Base voltage per part is set in mfg.
- Adjusted on the fly by PCU between 0.7-1.2V

VccSA supplies MC, CIO, IIO (yellow)

- Fixed voltage set in mfg. between 0.7-1.1V per part

VccP supplies IO's (blue)

- Fixed at 1.0V for all parts

VccD supplies DDR IOs (magenta)

- Fixed voltage set by the system at 1.5 or 1.35V

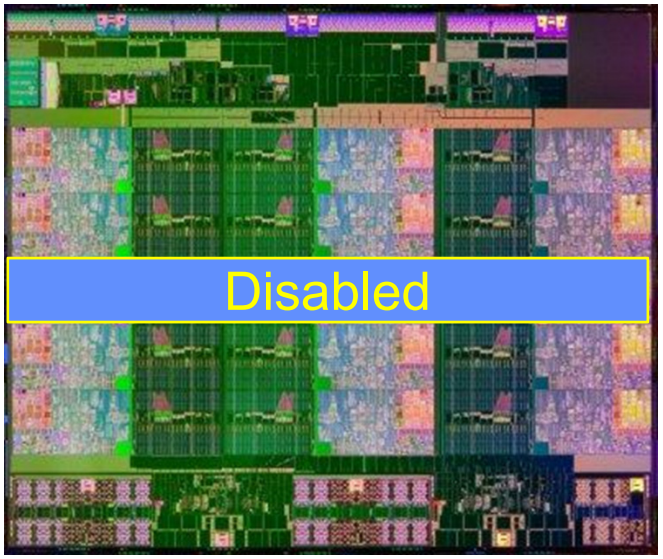
VccSFR supplies PLL's and DTS's (black)

- Fixed voltage set at 1.7V for all parts

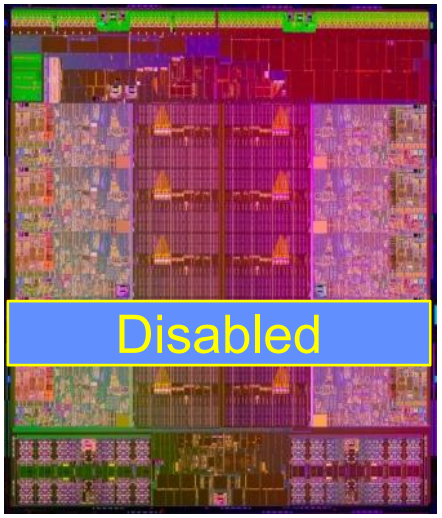
VccF (Green)

- Tied to package pin, shorted to VccP at socket

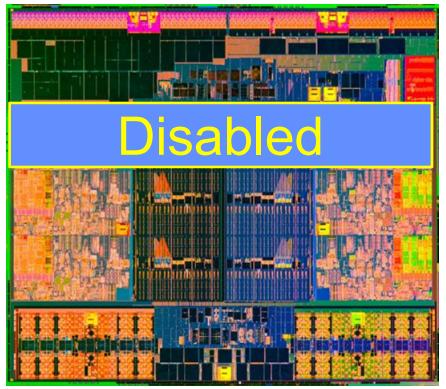
Core and Cache Recovery



15 ➔ 12 cores



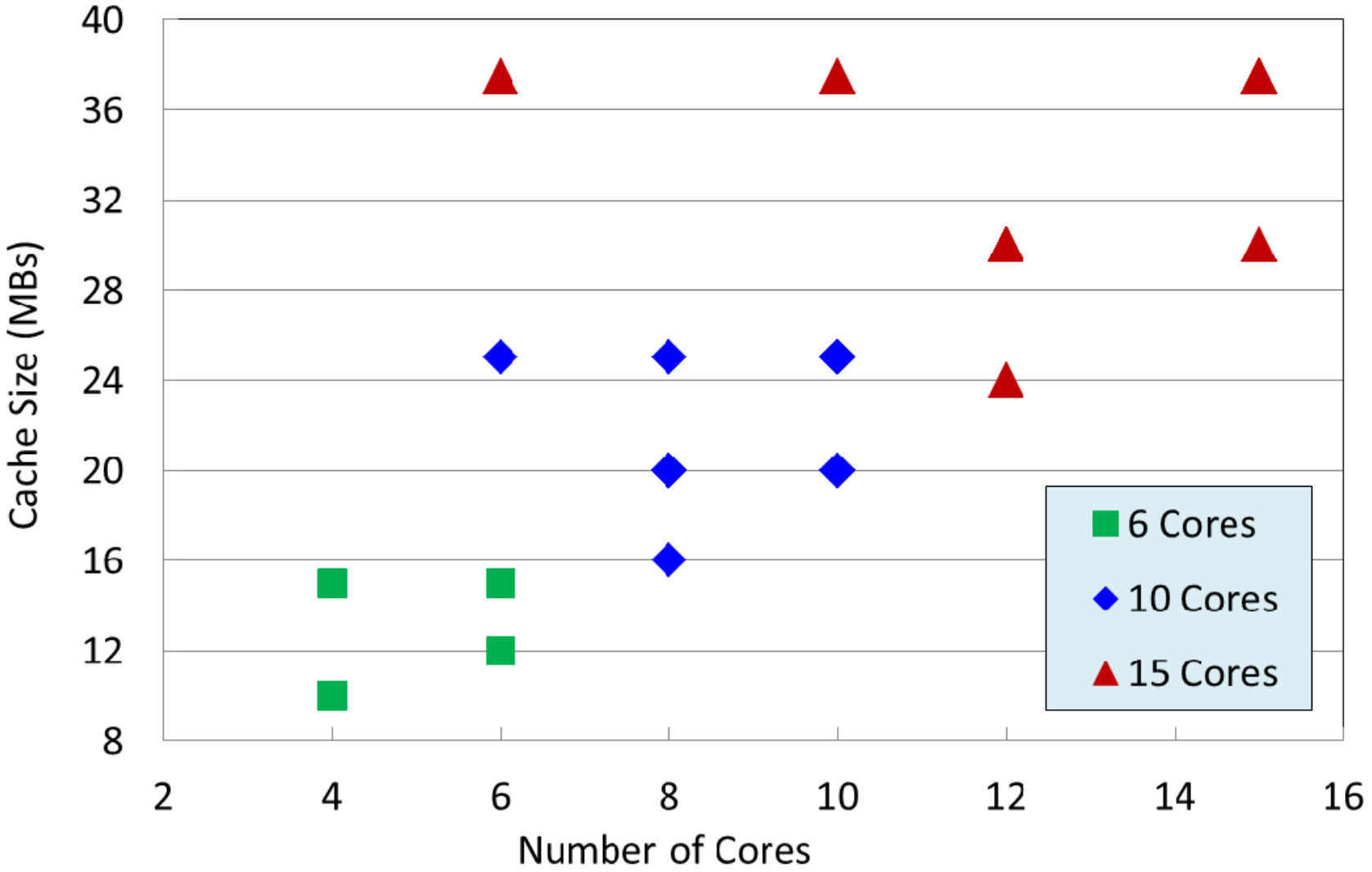
10 ➔ 8 cores



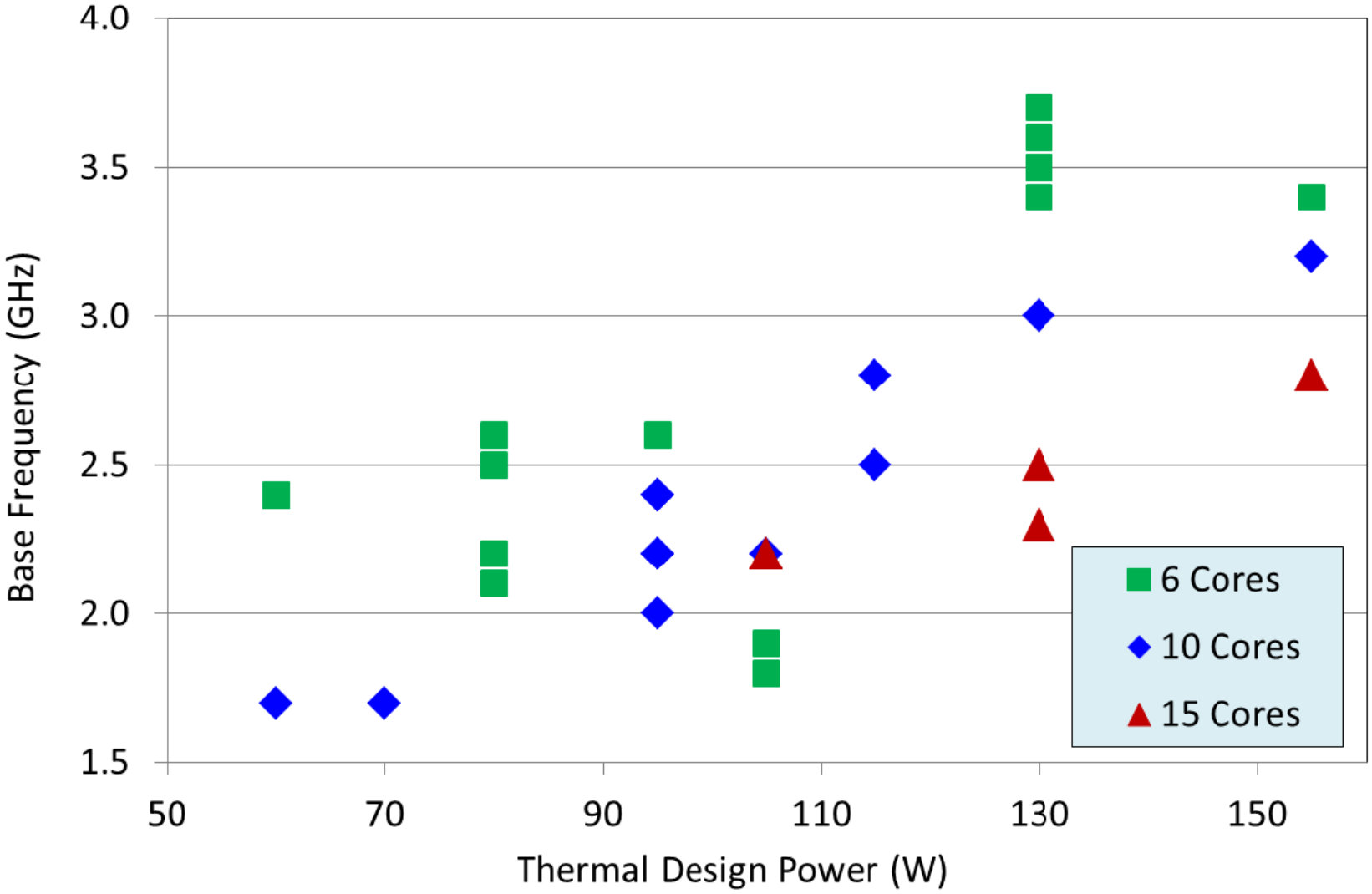
6 ➔ 4 cores

- Clock and power-gate any row in each die
- Cores and caches can be disabled separately
- Disabled cache slices in shut-off mode

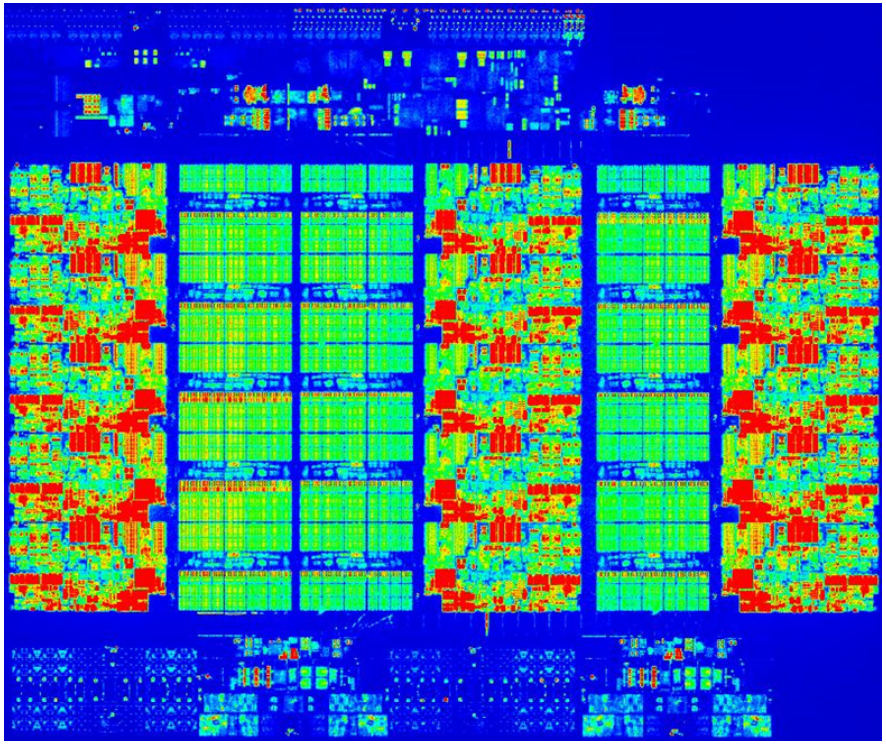
Core and Cache Sizes for Ivytown SKUs



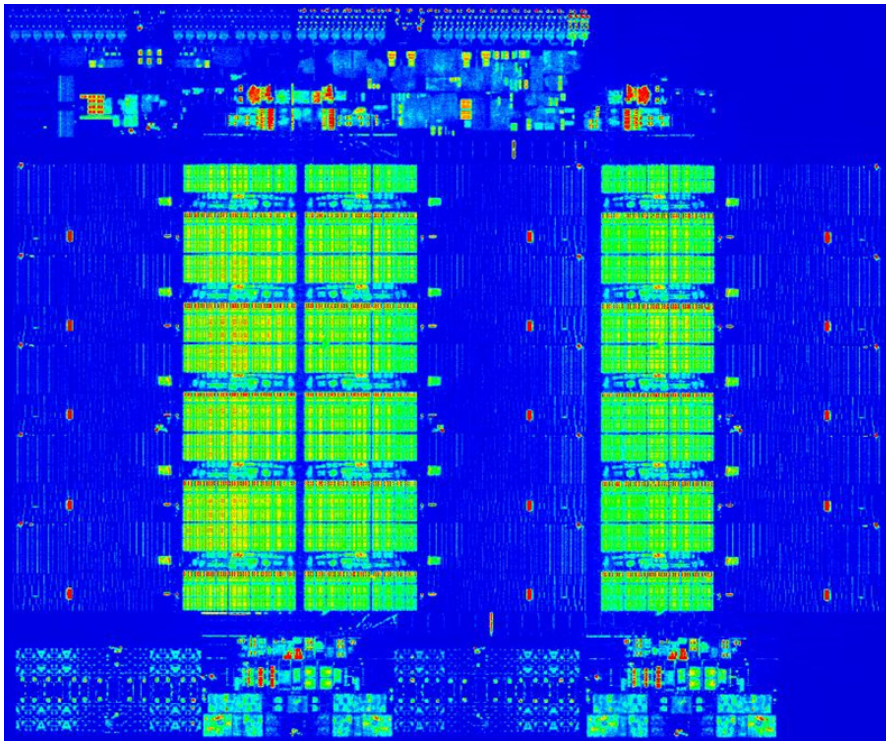
Frequency vs. Power for Ivytown SKUs



Idle Power State

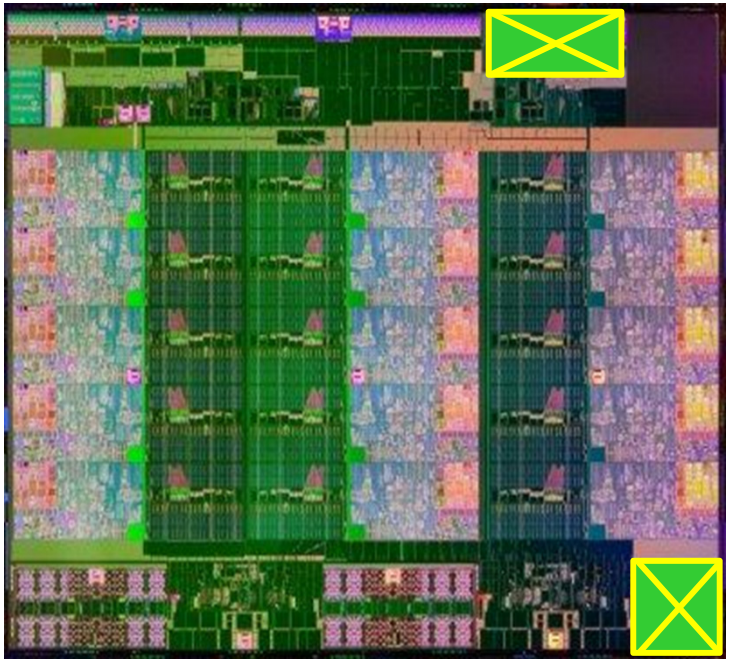


All cores and cache slices are enabled

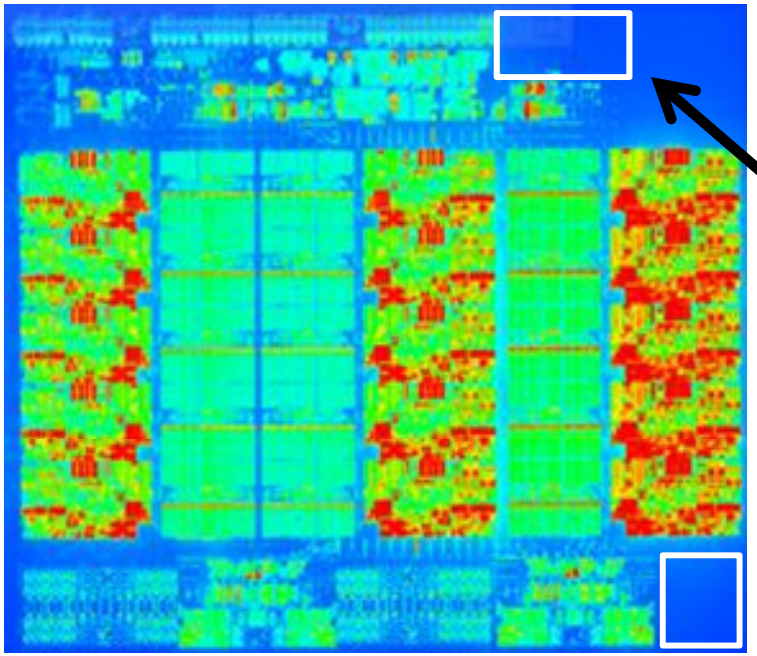


Cores are in C6 state using power gates in weak mode
Cache maintains state to be able to resume with low latency

IO Recovery



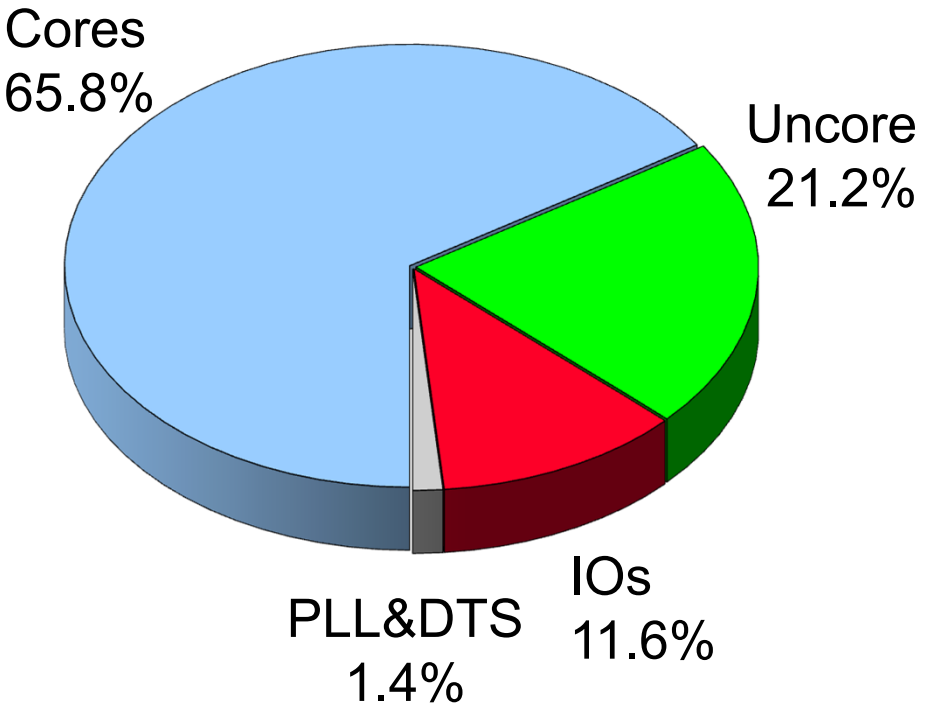
15c die in EP package

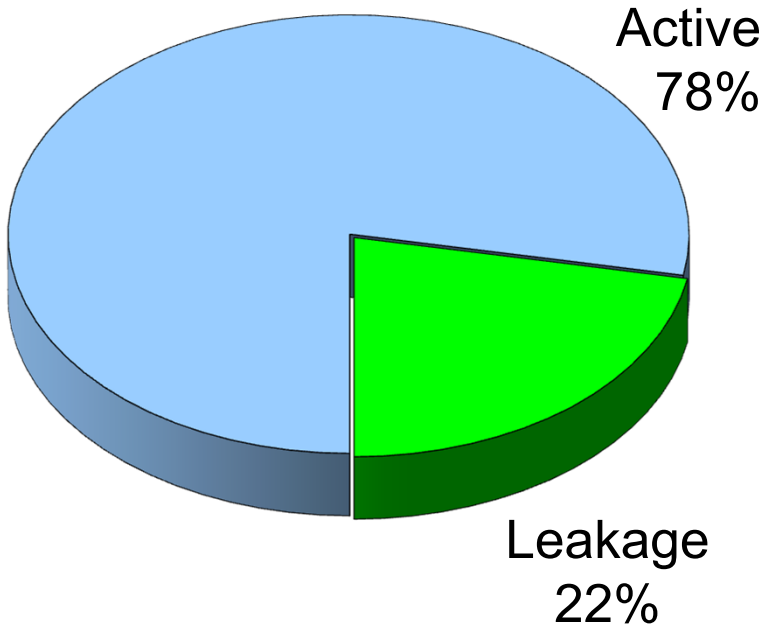


Infrared Emission Microscope View

- The 15-core die in EP package does not use the top right QPI port and bottom right VMSE port – they are grounded in the package
- Small yield recovery benefit, ~0.7W leakage savings

Power Breakdown





Leakage reduction → Long channel device usage: 63% cores, 92% uncore and IOs

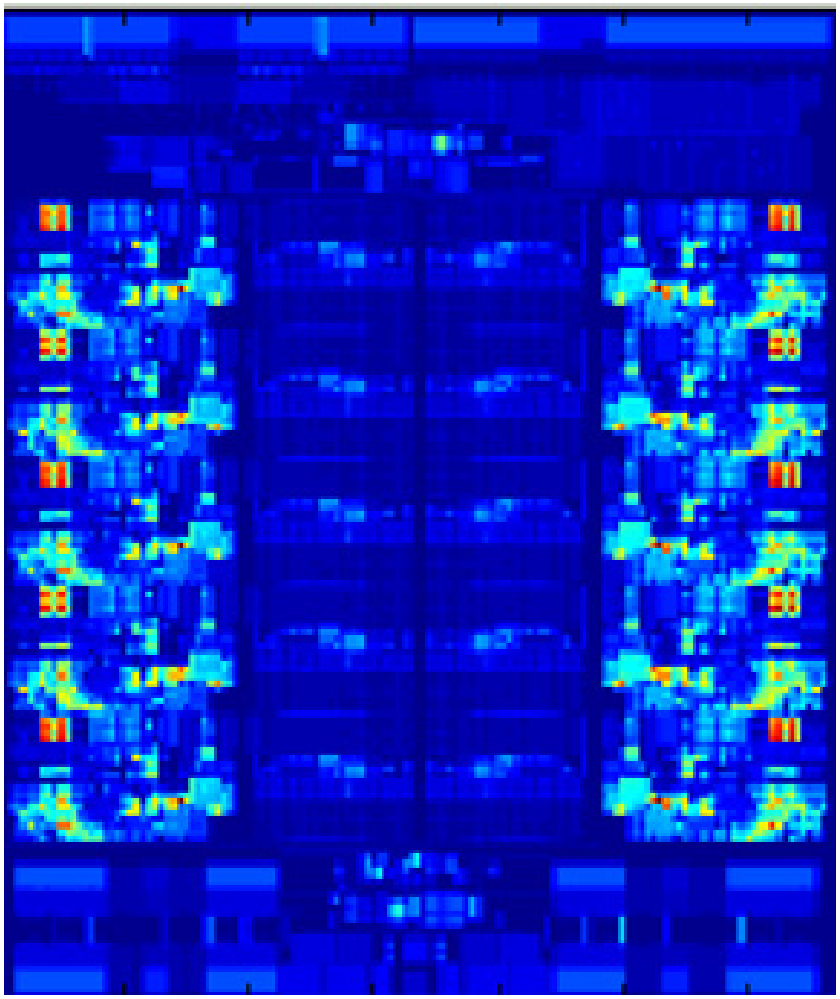
Power Reduction Techniques

- Fine granularity clock gating across die (Core/Uncore/IO)
  - Idle and targeted application power savings
  - Per X4 lane of PCIe, per QPI lane and per DDR channel
- Unused IO logic disable through package level short
  - Reduction in leakage → Helps both TDP and Idle power
- Configuration based clock gating on IOs → BIOS/PM

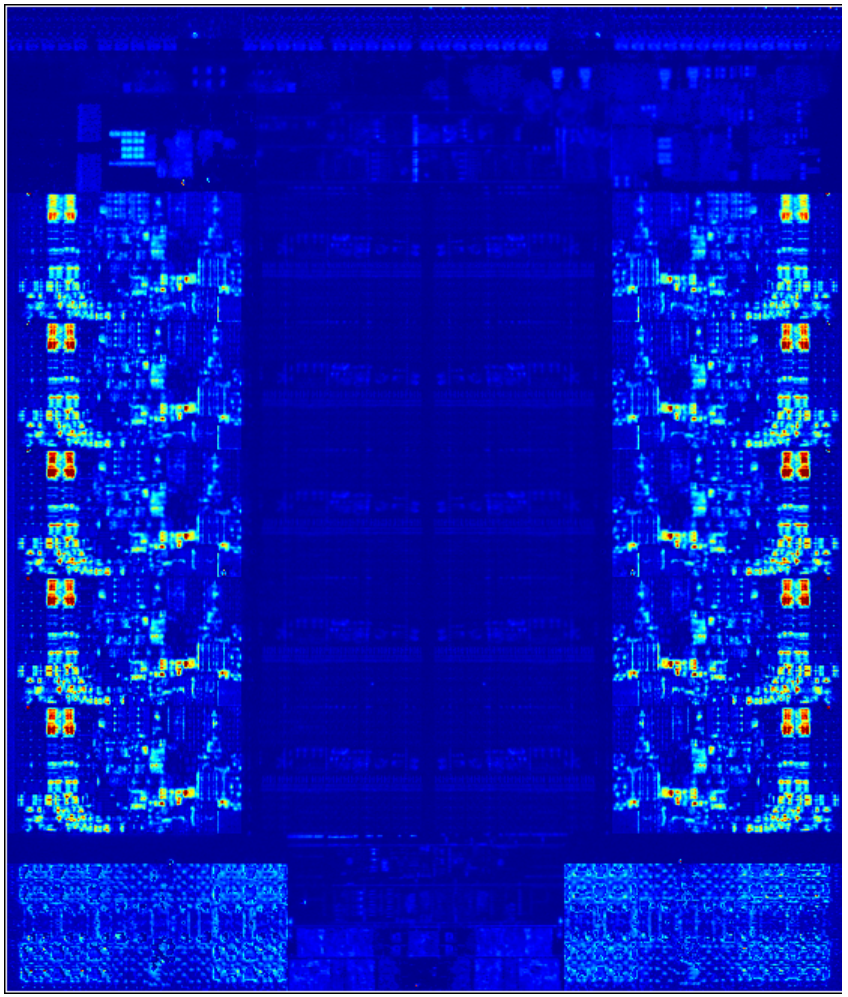
IO Status	IO Condition
Disabled IO	IO exists on die but not supported in a SKU
Unrouted IO	IO exists and is supported, but OEM chose not to route it
Unpopulated IO	The IO port is wired but not populated at boot but could be populated later
PC6 Idle	System is in Package C6 (PC6) idle, IO is in the lowest power state

- Reduction in redundant LLC activity based on resource
- Multiple HVM techniques for per part leakage/Cdyn tuning

10c Power Maps

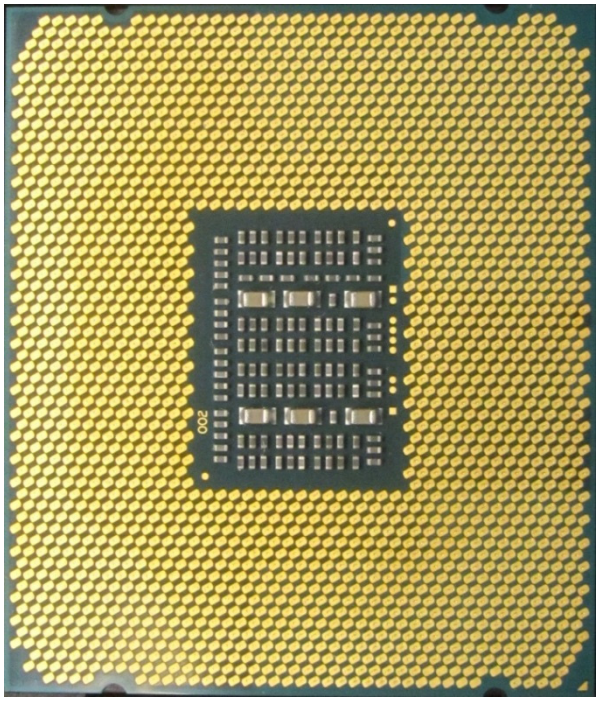


Pre-silicon power map



Silicon-based power map

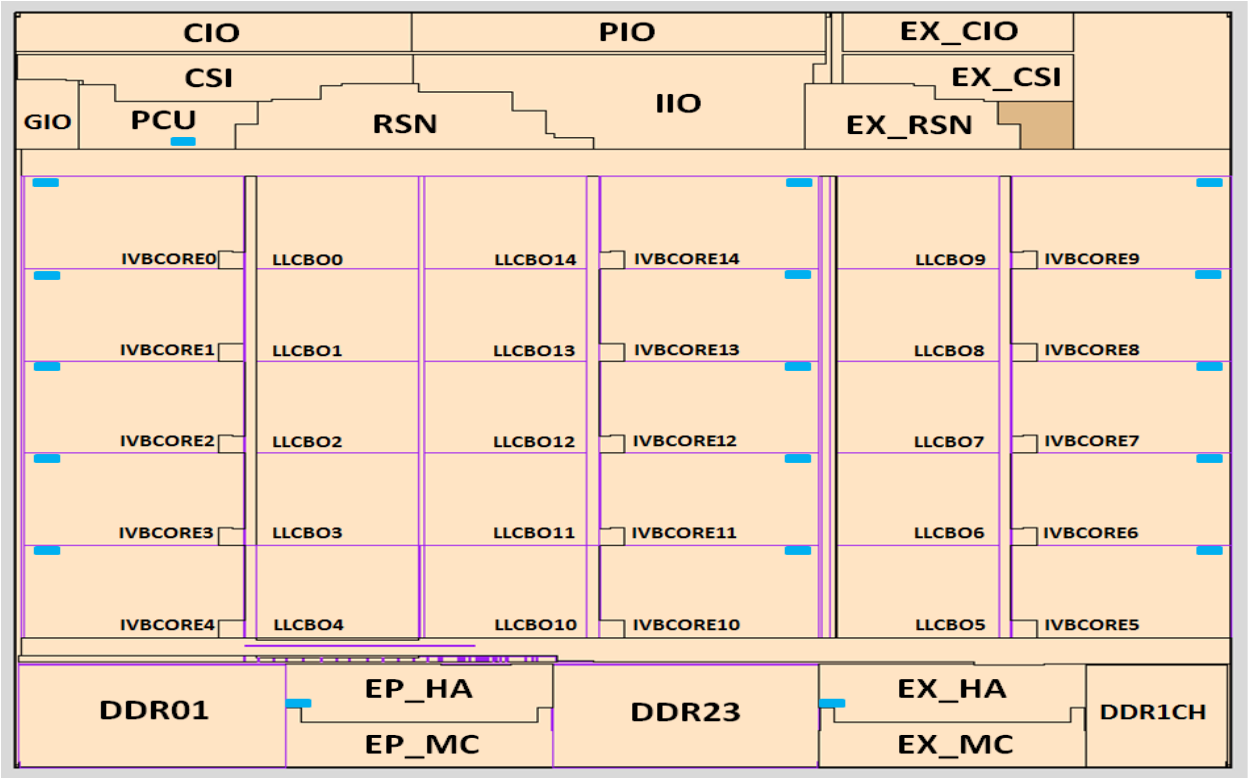
Package Details



- 20 layer organic substrate
  - 45.0 x 52.5 mm
  - 6-8-6 layer stacking
- Integrated heat spreader
  - 49.0 x 37.0 mm
- 100% lead-free and halogen-free

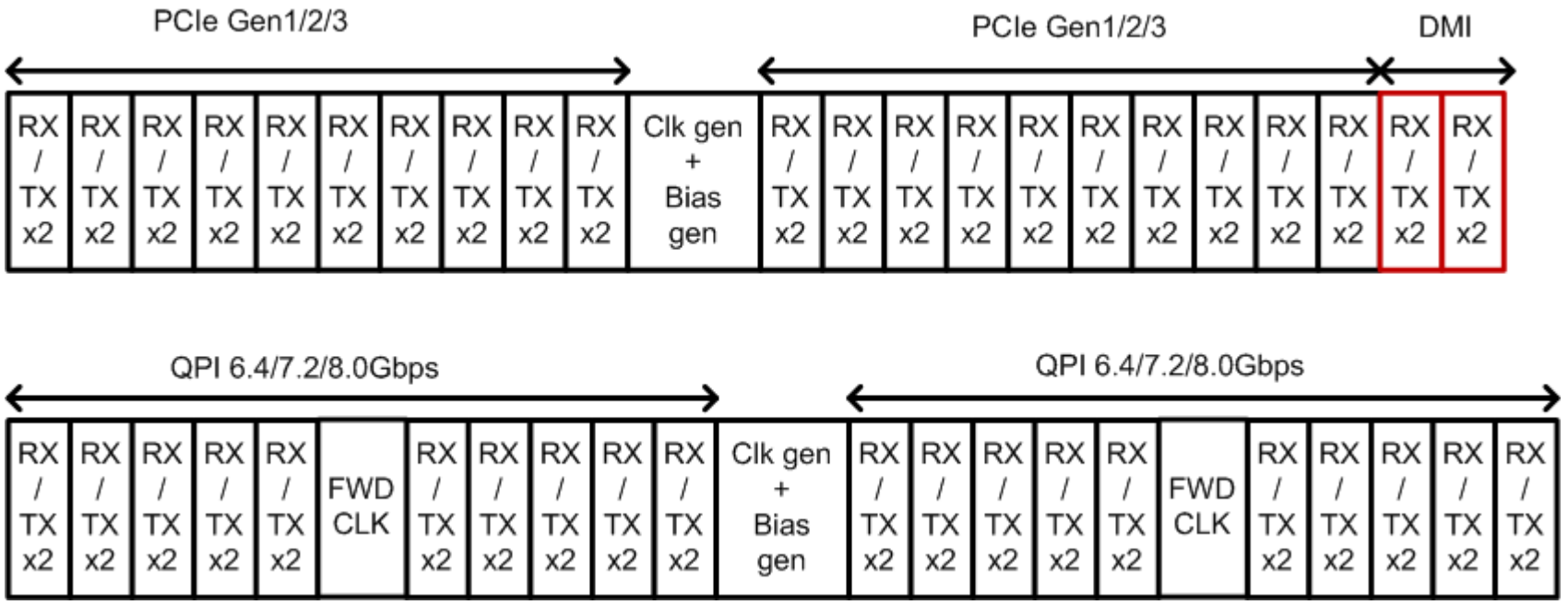
- 2011 total lands at 1.016 mm hexagonal pitch
  - 984 signal IO's
- 16 x 22 mm cavity
  - Decoupling capacitors on package bottom directly opposite circuits

Temperature Sensors



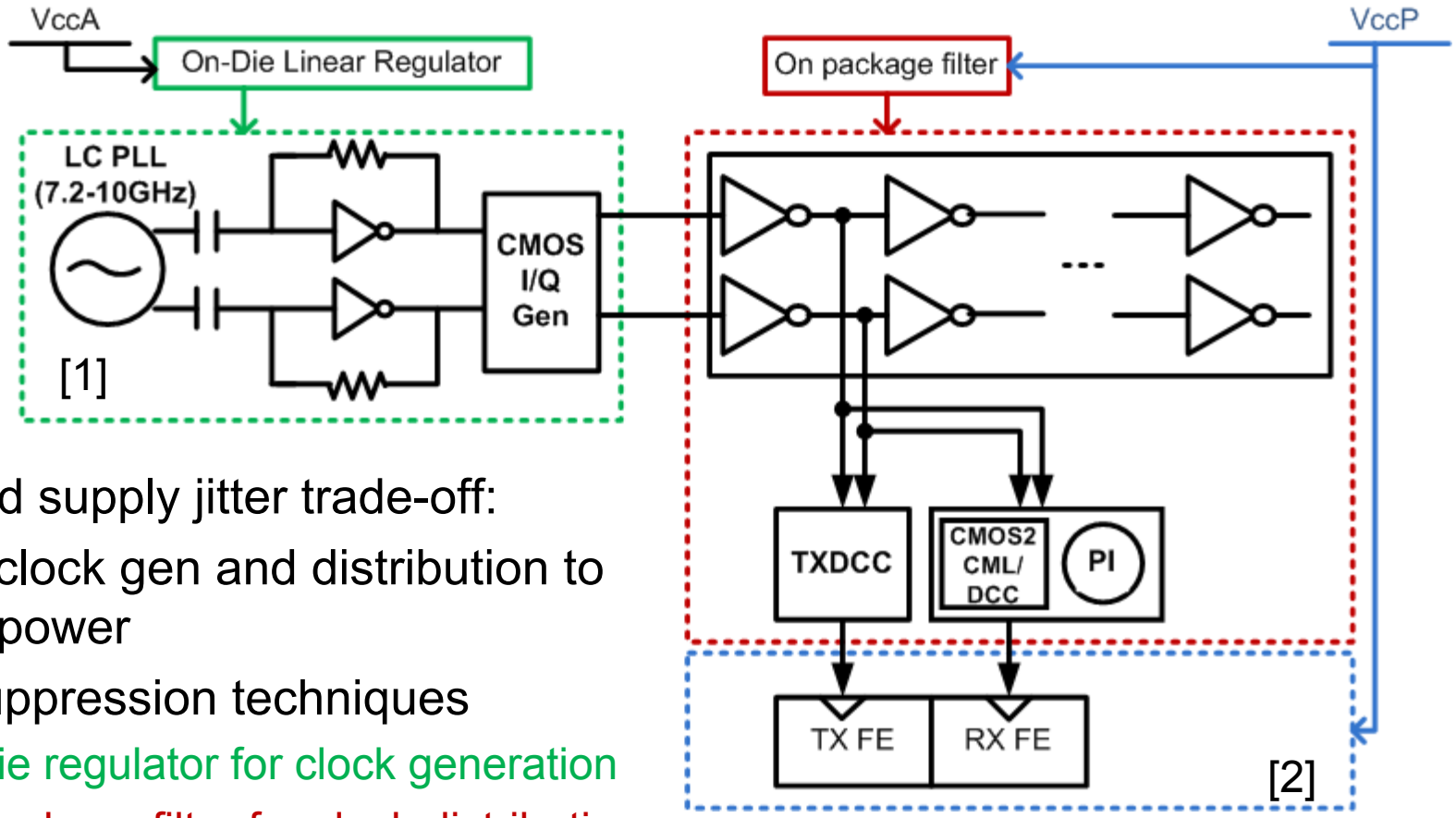
- Used for Thermal throttle @Tjmax and Thermtrip for catastrophic thermal event (e.g., a system fan failure)
  - One in each core near hot spot
  - Three in Uncore area of the die
- Temperature data available through PECI bus for system fan control

High-Speed Serial Links



- 40 lanes PCIe Gen1/2/3 at 2.5/5.0/8.0 Gb/s
- 4 lanes Direct Media Interface (DMI) at 2.5/5.0 Gb/s
- 40 lanes Quick Path Interconnect® (QPI) at 6.4/7.2/8.0 Gb/s
  - Forwarded clock for better jitter tracking and lower BER in QPI
  - Additional 20 QPI lanes for EX segment

Link Performance Optimization



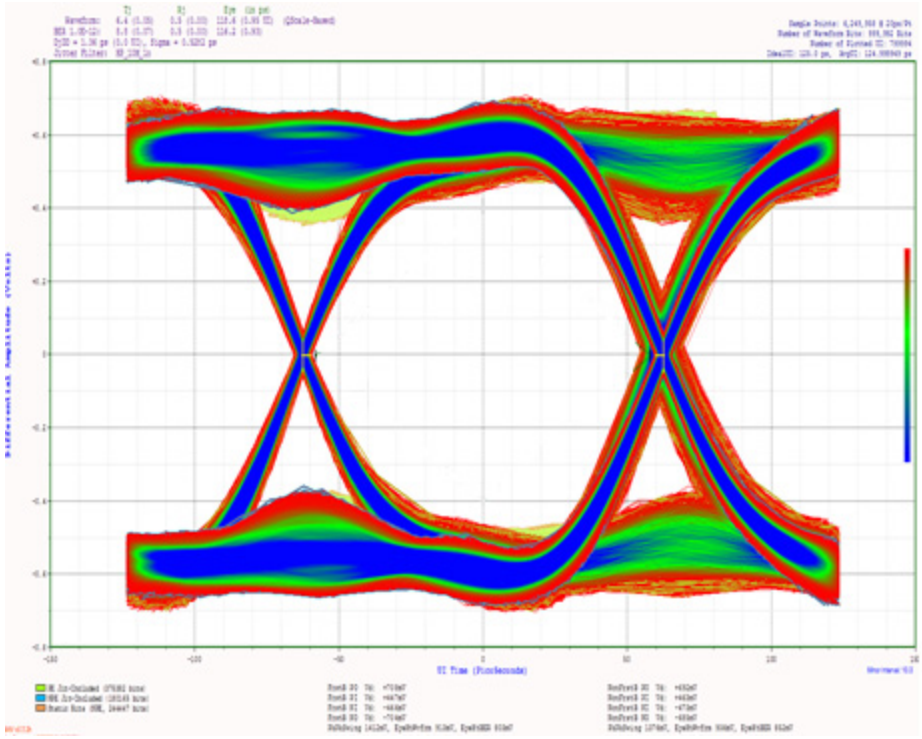
Power and supply jitter trade-off:

- CMOS clock gen and distribution to reduce power
- Jitter suppression techniques
  - On die regulator for clock generation
  - On package filter for clock distribution (including phase interpolator)
  - Lane-staggering and on-die MIM decap for VccP noise reduction

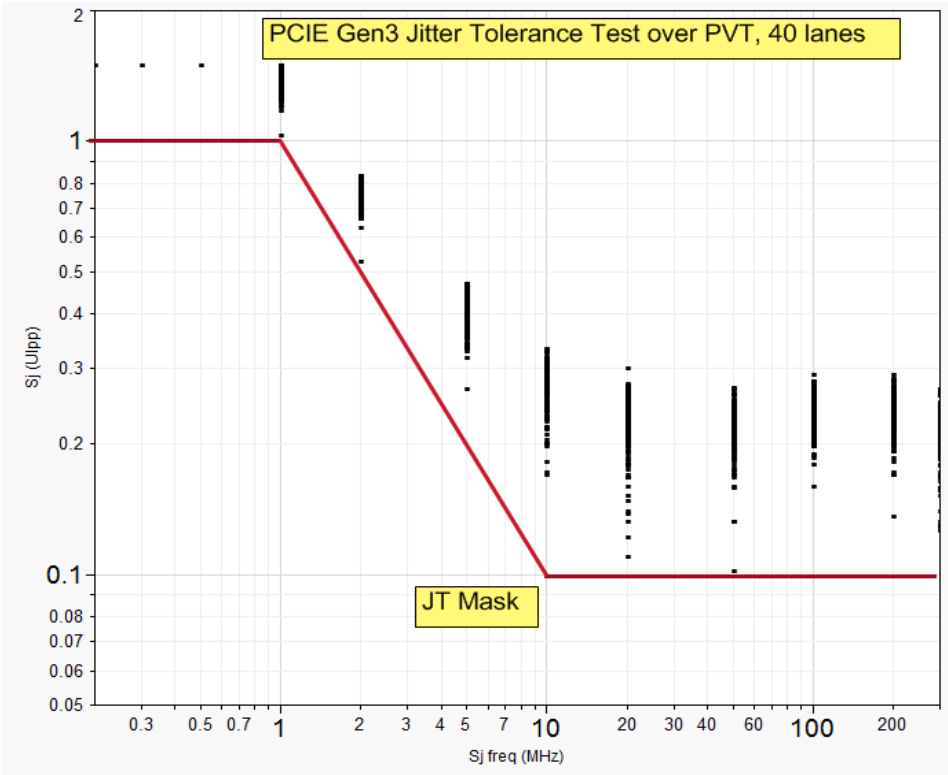
[1] S. Li, et al, Clock generation for a 32nm server processor with scalable cores. ISSCC 2011
[2] F. Spagna, et al, A 78mW 11.8Gb/s serial link transceiver with adaptive RX equalization and baud-rate CDR in 32nm CMOS, ISSCC 2010

Serial IO Experimental Results

TX Eye Diagram



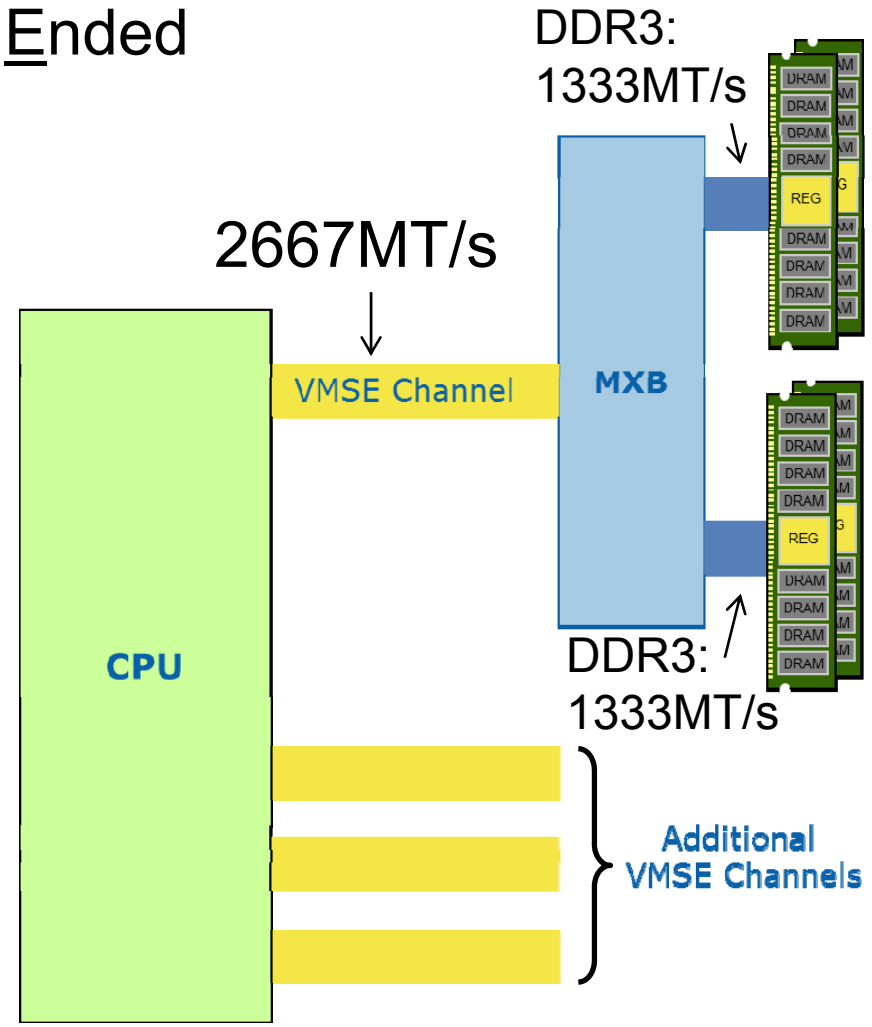
RX Jitter Tolerance Tests



- Circuit optimized and met PCIe 3.0 Base-spec and PCIe 3.0 Card Electromechanical (CEM) Spec over PVT conditions and server platform design guidelines:
  - RX AFE Gain/Peaking/Bandwidth optimization
  - RX DFE summer and Data/Error sampler offset and mismatch reduction
  - Residue ISI reduction and RX/TX Clock quality (Dj/Rj/DCC)

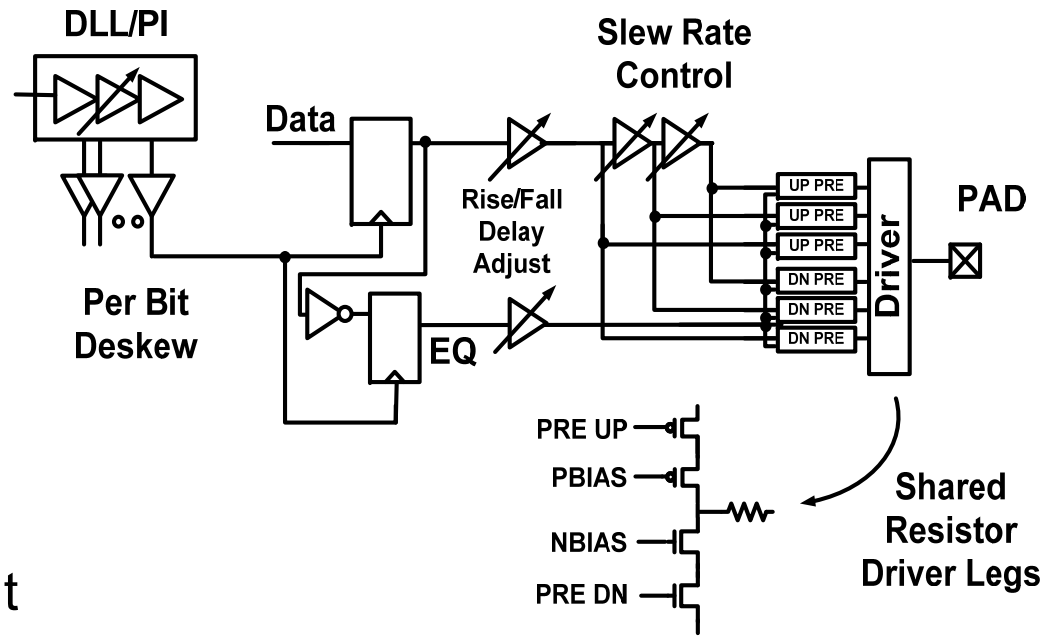
VMSE System Architecture

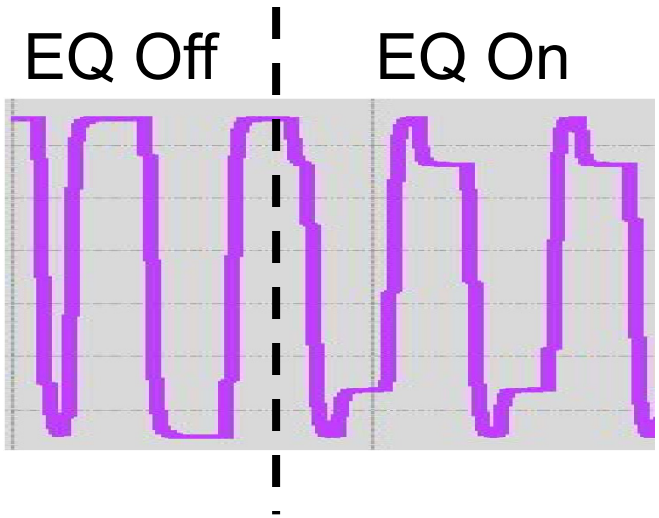
- VMSE = Voltage Mode Single Ended
  - 72 pins (8 data + 1 ECC bytes)
- Bi-directional interface between CPU and Memory Extension Buffer (MXB)
  - Each buffer has one VMSE channel on front-side and two DDR channels on back-side
- VMSE runs on the same pins as DDR3 for common die
- ~75GB/s 4-ch effective BW at 2667MT/s per lane



DDR3 / VMSE TX

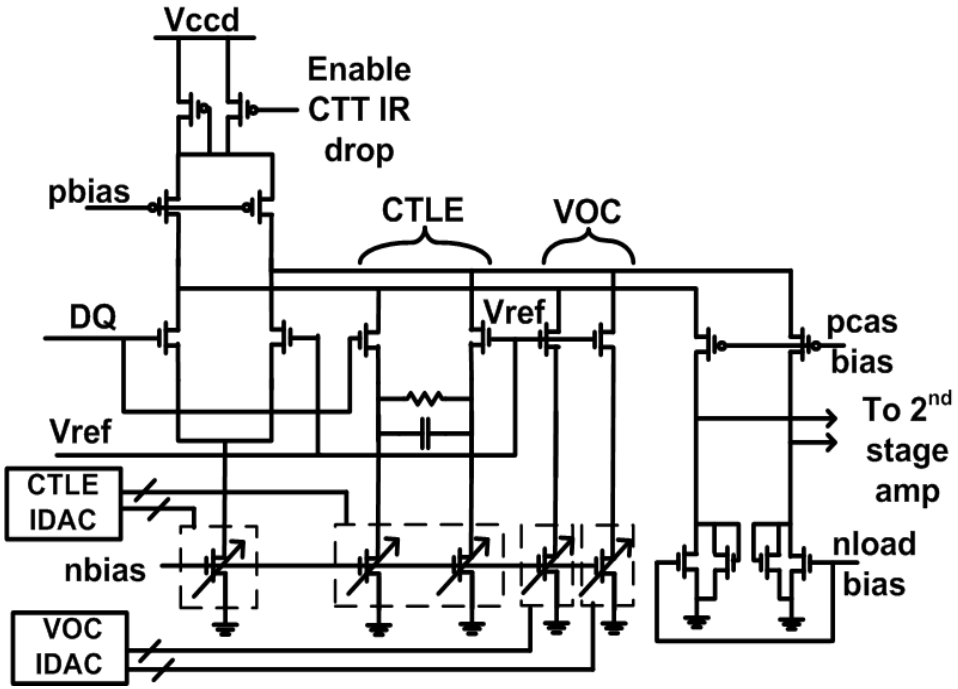
- 800-2667MT/s range for combo buffer
- Stacked thin gate output stage for EOS protection
- Shared resistor for area and Cpad savings
  - Minimal shoot-through current
- 2-tap EQ with programmable coefficient
- Adjustable rise/fall delay and slew rate control

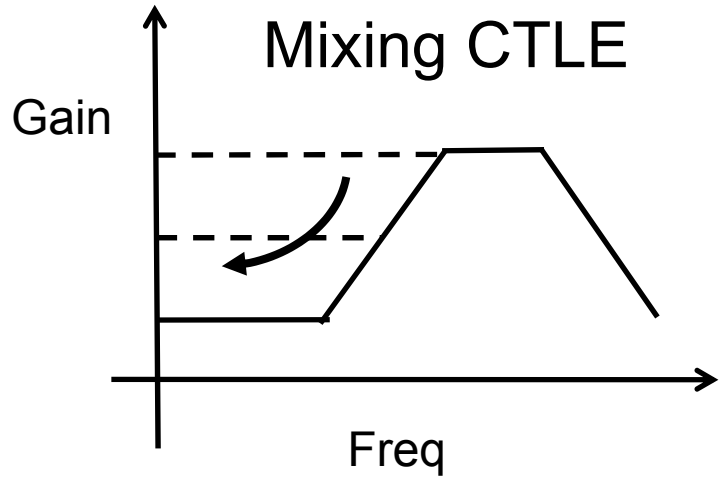




DDR3 / VMSE RX

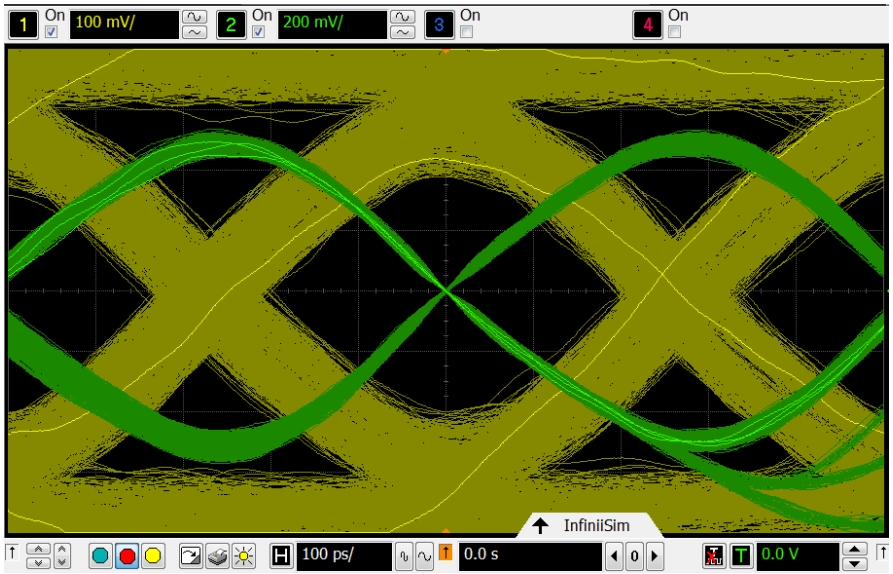
- DDR3 center-tap and VMSE Vddq terminated inputs
- Folded architecture for max headroom
  - IR drop switch for VMSE vs DDR3 common-mode
- Triple diff-pair input stage
  - Main input
  - CTLE for Rx EQ
  - VOC for offset cancellation
- CTLE coefficient control via mixing 2 diff-pairs to minimize RC parasitics on virtual Gnd





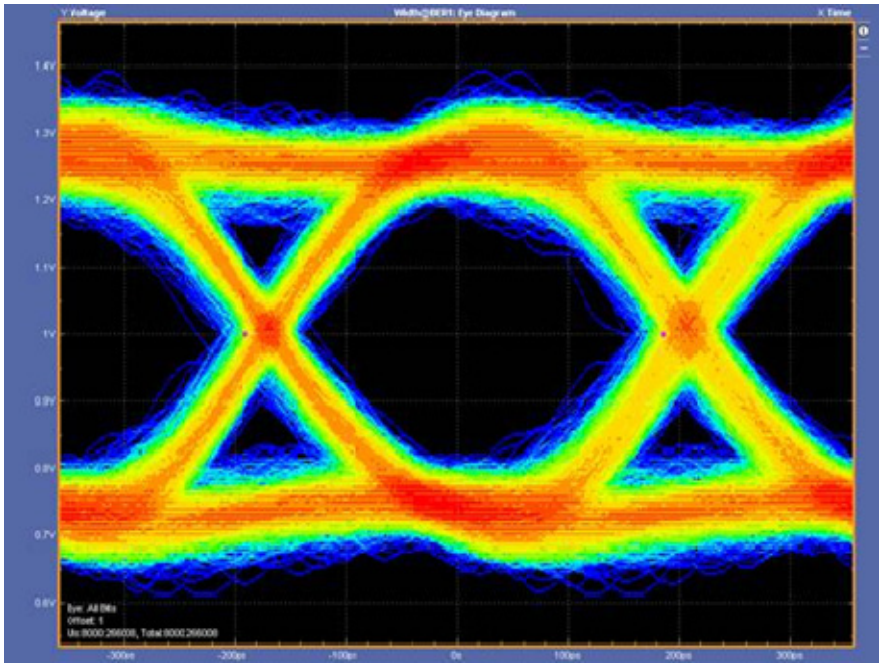
DDR3 and VMSE Eye Diagrams

DDR3 1867Mb/s Write



DDR3 Setup: 2 DIMM per channel, 4-rank, measured with DRAM device interposer

VMSE 2667Mb/s Write



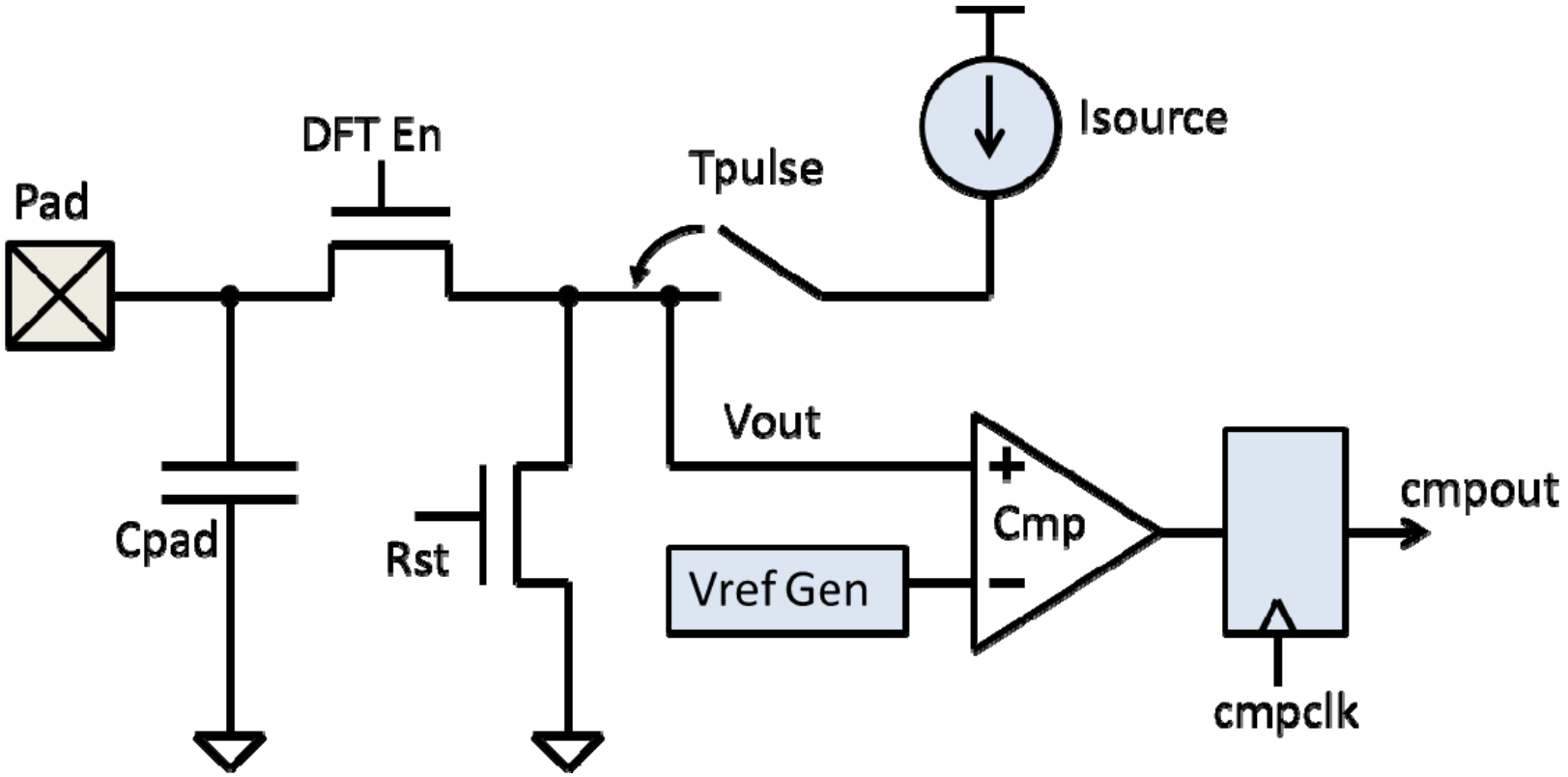
VMSE Setup: 14” link with CPU socket and 1 connector, measured at PTH via on backside of MXB

Physical Design

- Ivytown processor design continued the trend for larger synthesized blocks [1]
- High performance designs of the order of 2 million gates flat were implemented using advanced multi-core enabled synthesis/static timing techniques
- Enhanced synthesis, extraction and verification technologies enabled equivalent turnaround time as previous generations with 2x increase in complexity
- Special design methodologies for signal integrity and circuit quality enabled minimal design convergence iterations

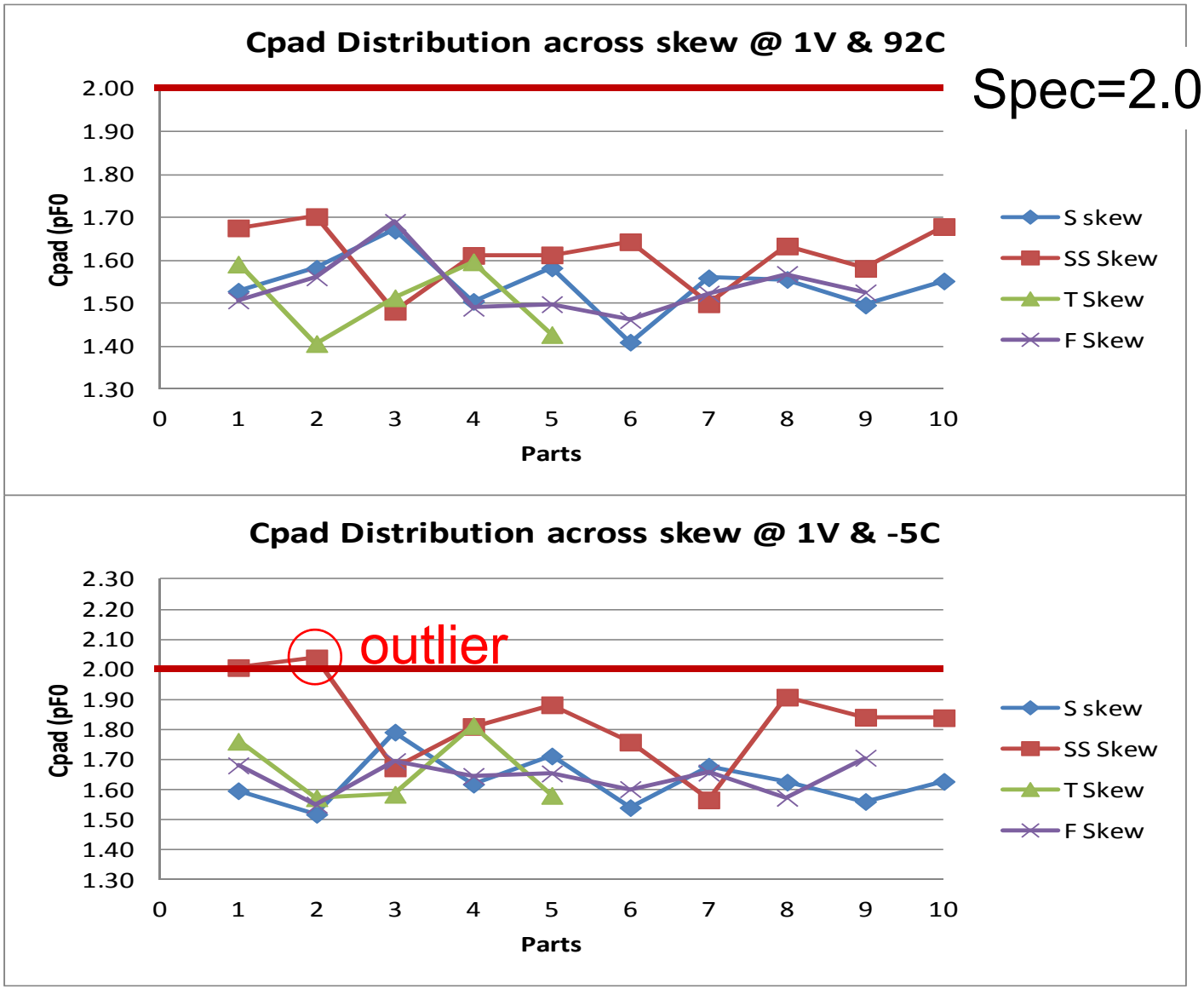
[1] R. Varada et al, Superblock: A Method for Synthesizing Large High Performance Designs without Hierarchy Limits, DAC 2010

Cpad Measurement Circuit



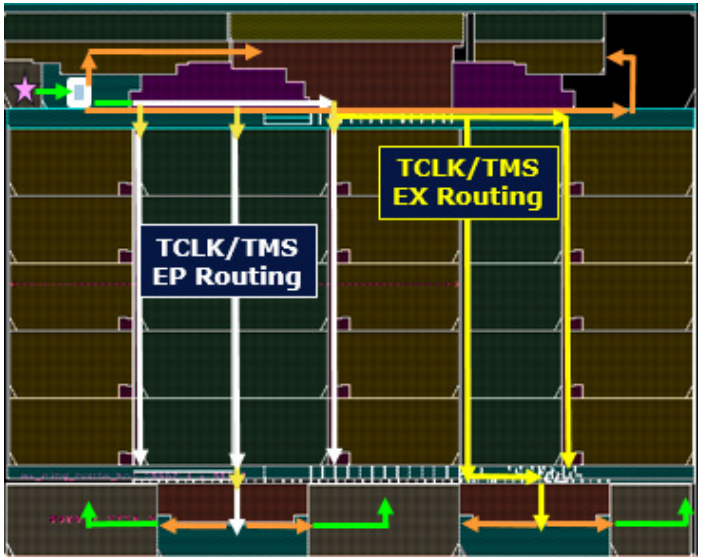
$$\text{Cpad} = \text{Isource} * \frac{\text{Tpulse}}{\text{Vout}}$$

DDR Cpad Result



DFT and DFM Features

- Large uncore area includes very high scan insertion rate
  - Better at-speed and stuck@ coverage, less SBFT dependency
- Test satellites are added in a tile fashion across die
  - In-die Variation Probes, Voltage Droop Monitors and Inducers
- IO : Digital and Analog IO loop back support
- Cache and small array: Programmable on die test engine
- Multiple TAP controllers:
  Help manipulate DFx features and logic through internal clock alignment
- Test clock distribution:
  Forwarded clock across die ►



Summary

- Presented an enterprise-class 22nm 15-cores, 30-threads Xeon® Processor with 37.5MB on-die shared L3 cache
  - Largest transistor count for a microprocessor
  - Modular floorplan supporting chops to 10 and 6 cores
  - Socket compatible with Romley platform (Socket-R)
  - Cache Vccmin improvement from per-die redundancy, per-die RWA settings and DECTED
- Active power and leakage reduction techniques
  - Multiple voltage and clock domains
  - Extensive use of long channel devices
- VMSE provides a 2667MT/s bi-directional interface between processor and Memory Extension Buffer (MXB)
- On-die Cpad measurement circuit ensures electrical spec compliance for memory and serial links

Acknowledgement

The authors gratefully acknowledge the work of the talented and dedicated Intel team that implemented this processor.

Please visit our demo this evening

AMD's "Steamroller" Core, an x86-64 Core Implemented in 28nm Bulk CMOS

Kevin Gillespie[1], Harry R. Fair III[1], Carson Henrion[2], Ravi Jotwani[3], Stephen Kosonocky[2], Robert S. Orefice[1], Donald A. Priore[1], Jonathan White[1], Kathryn Wilcox[1]

[1] AMD, Boxborough, MA

[2] AMD, Fort Collins, CO

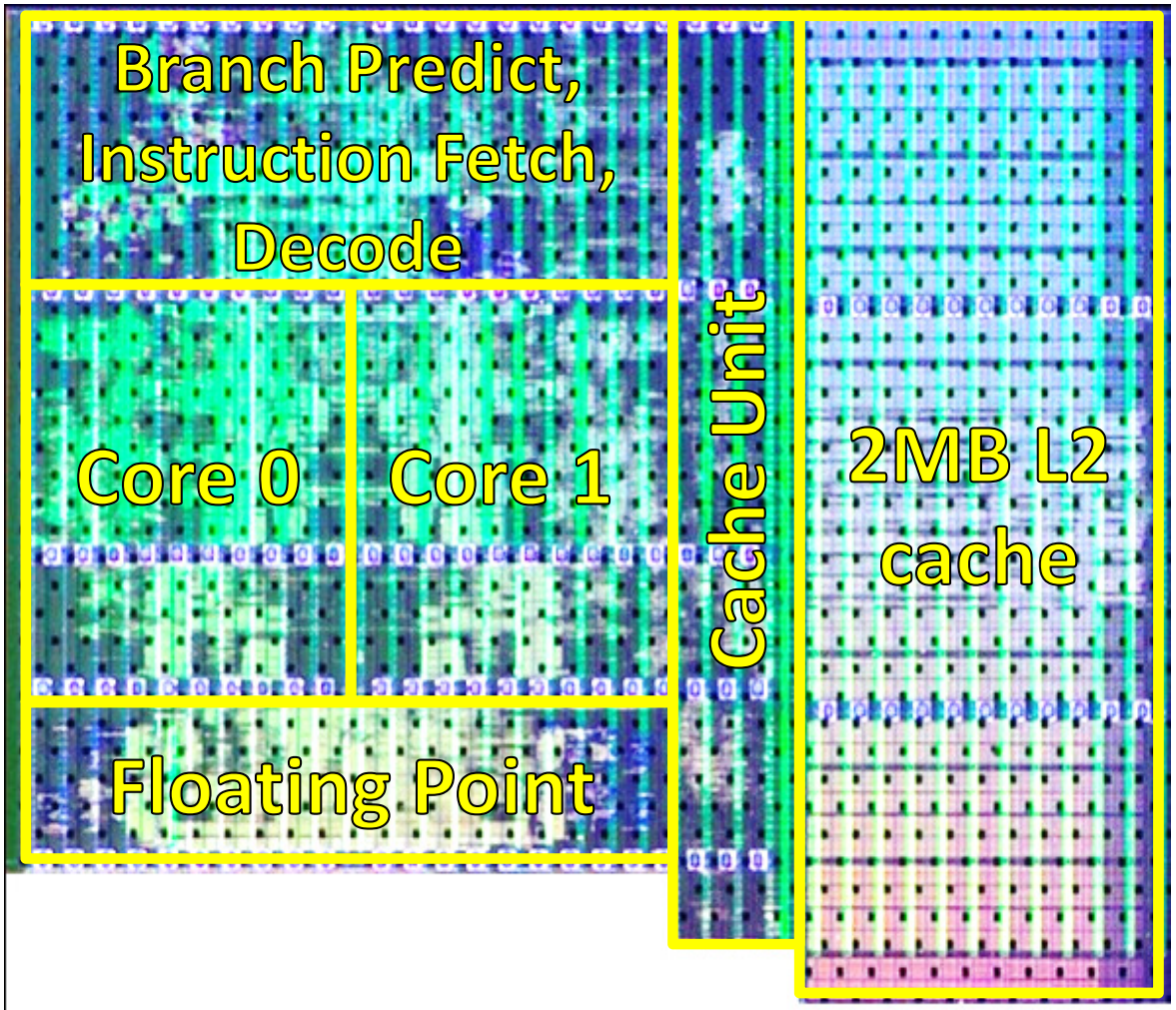
[3] AMD, Austin, TX

Outline

- Module Overview
- Power
- Resonant Clocking
- Distributed Power Gating
- L2 Power Gating
- Power Supply Monitors
- Soft Error Improvement
- Conclusion

Steamroller Module

- 28nm Bulk CMOS
- 12 metal layers
- 0.7v – 1.35v
- Area = 29.47mm^2
  - 2MB L2: 10.86mm^2
- 236M transistors
- IPC = +14.5%
- +500MHz
- -38% power @ 3GHz
- ~437K flops
- 63 unique macros



Technology Challenges

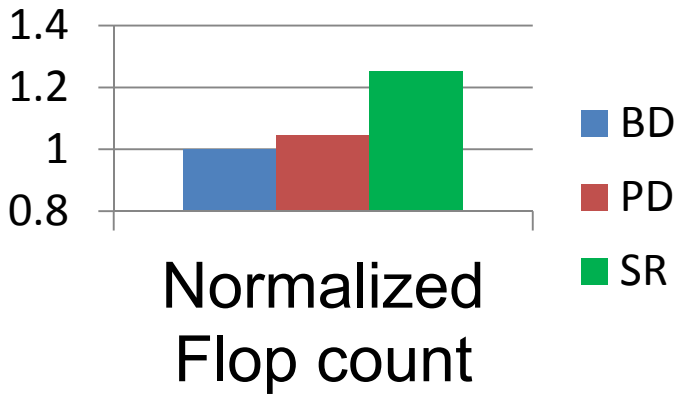
- Metallization changes
  - Loss of 1.25x layers
  - Increased RC
- ~10x Increase in Soft Error Rate
- Higher source/drain cap
- Well tap area bloat
- Sub-threshold leakage temperature dependence

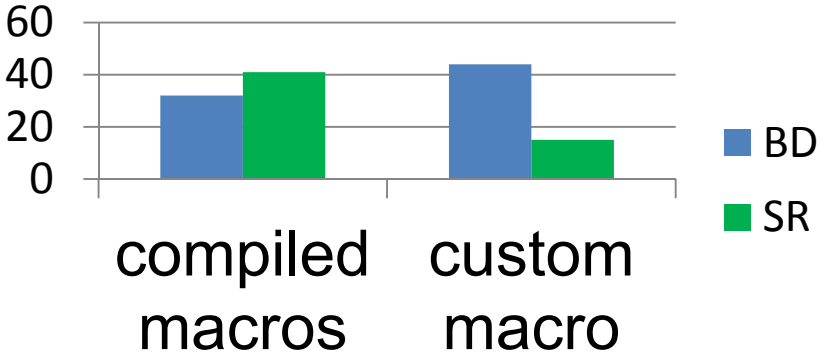
Metal Stack	32nm SOI	28nm Bulk
1x	3	6
1.25x	2	0
2x	3	3
4x	1	1
16x	2	2
Total	**11**	**12**

Major Structure Comparison

	Steamroller	Piledriver
L1 Instruction Cache	96KB 3-way	64KB 2-way
Instruction Decoder	1 per Cluster	1 Shared for 2 Clusters
FPU	3 pipe, same # of Ex units	4 pipe
L2 Sizing	Dynamic w/ power gating	Static
Branch Target Buffer	10K	5K
uOp Dispatch Queue entries	40	32
Int Physical Register File	112	96
Int Scheduler entries	48	40
FP Physical Register File	176	160
Load Queue entries	48	44
Store Queue entries	32	24
Probe Buffer entries	12	8

Physical Statistics

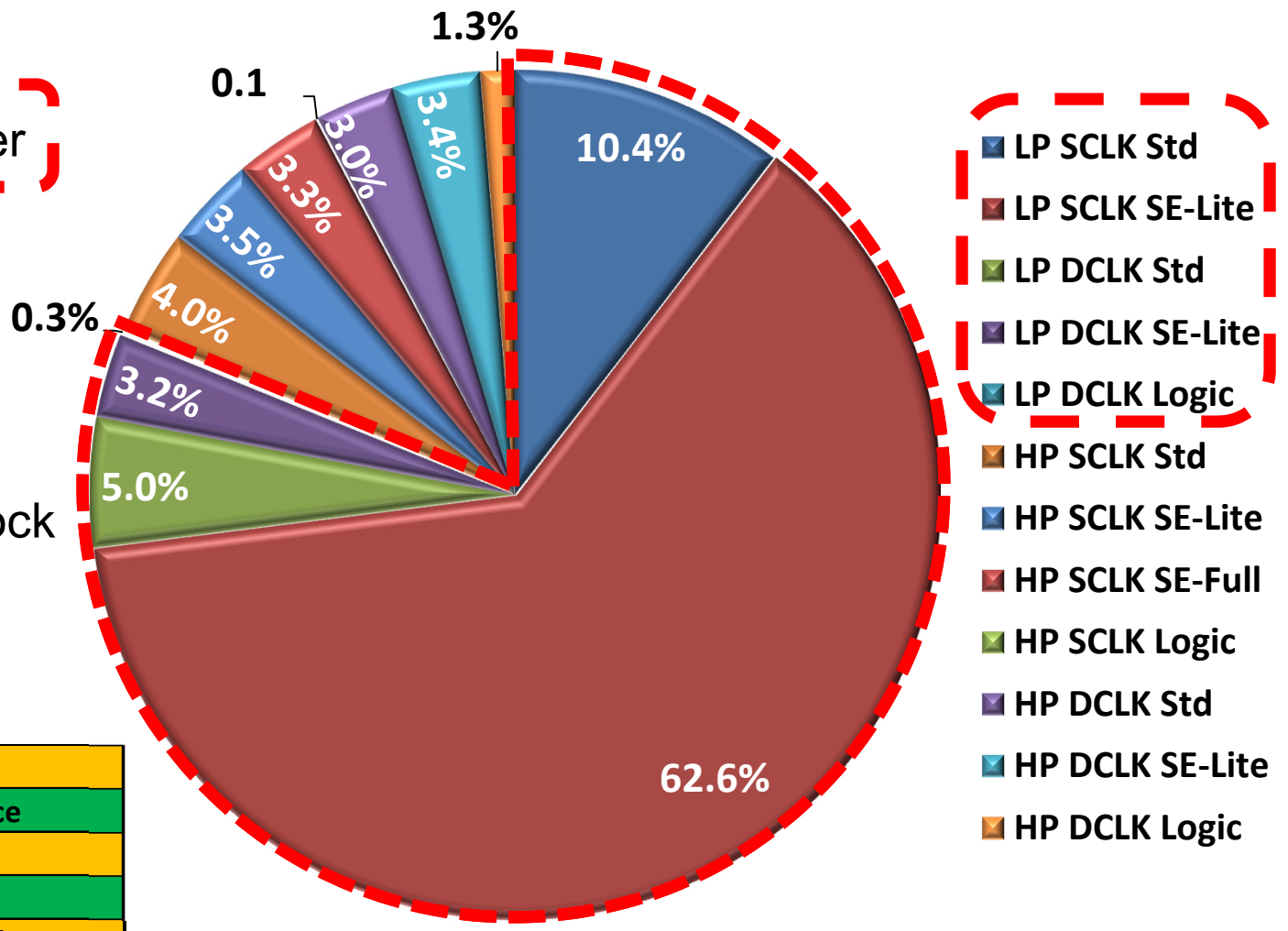




- Progressive generations migrating to higher percentages of SAPR
- Flop counts increased due to converting custom structures to flop arrays as well as architectural improvements
- Custom structures reduced to improve efficiency

Flop Statistics

- 81.4% Low Power



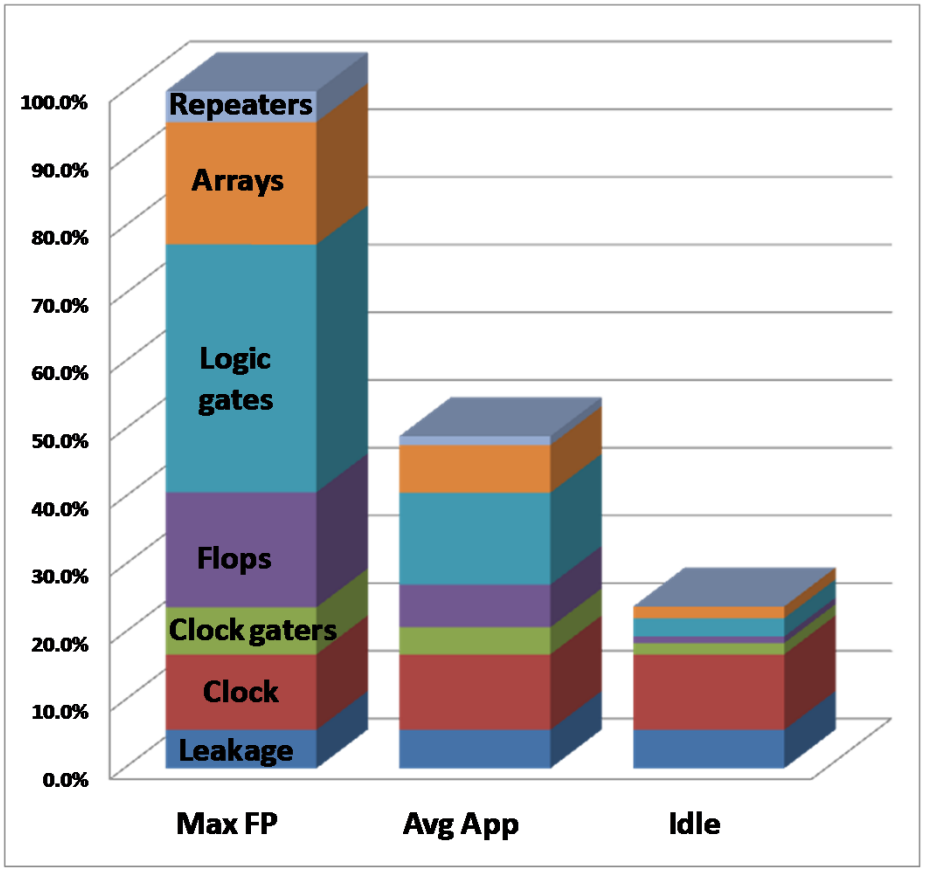
- 83.9% Single Clock
- 69.4% Soft Error

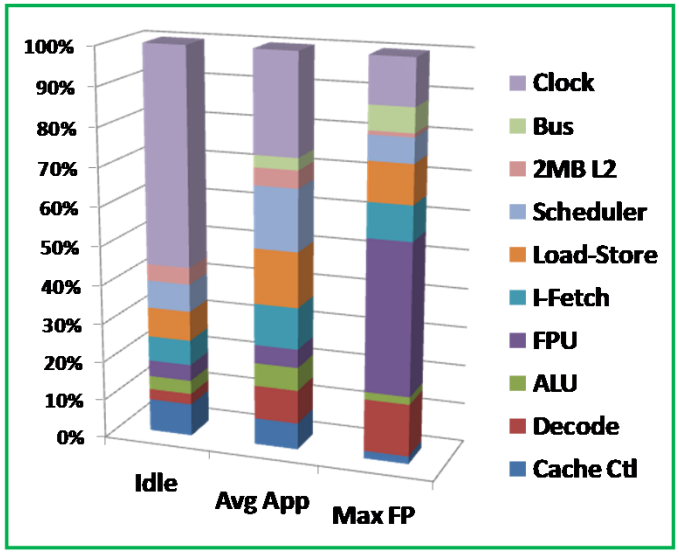
LP	Low Power
HP	High Performance
SCLK	Single Clock
DCLK	Dual Clock
SE-Lite	Improved Soft Error Robustness
SE-Full	Soft Error Tolerant
Logic	Logic Front-End

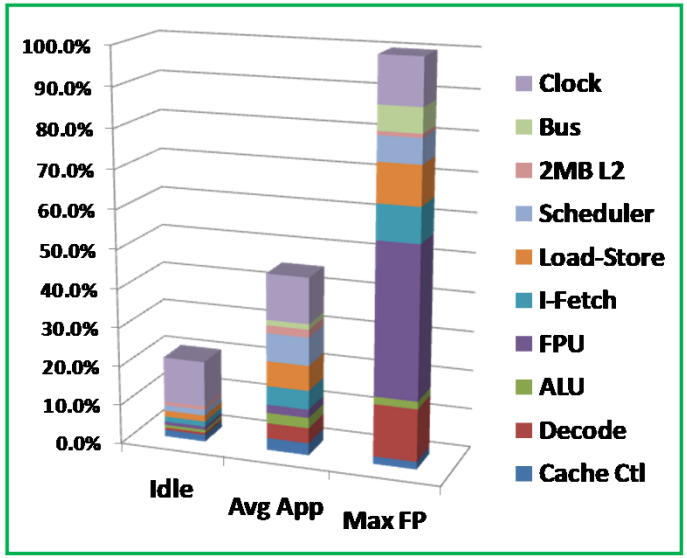
Power Improvements

- 10% Cac reduction through design
- Early and continuous tracking of flop activity in rtl
- Targeted improvements
  - L2 cache resonant clocking
  - L2 cache way power gating
  - Increased use of single clock soft edge flop
  - Redesign of targeted high power blocks
  - Enabled power-aware SAPR features
- Replaced internally developed power measurement flow with industry standard
- Design of Experiments approach to investigate power reduction options

Active Power Breakdown







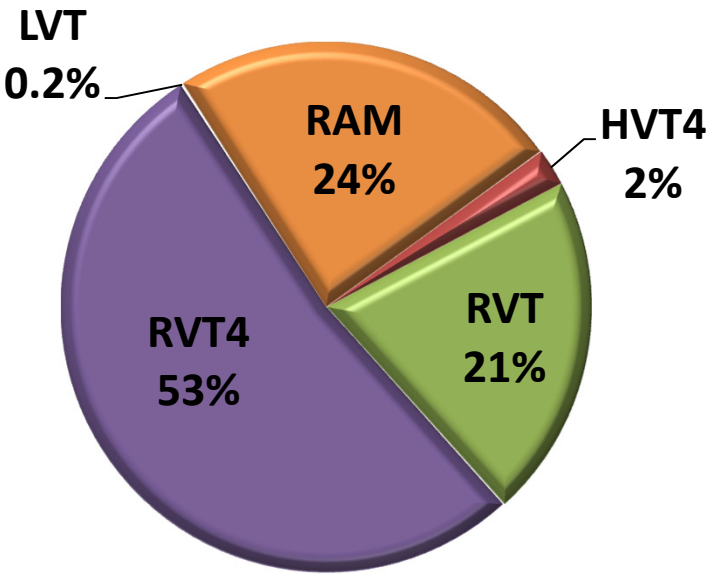
- Substantial power reduction across applications

Leakage Breakdown

- Low leakage
  - Vt mix
  - Device targeting

Normalized Device Leakage	32nm	28nm
HVT	1.00	0.53
RVT	5.33	3.67
LVT	16.67	19.80

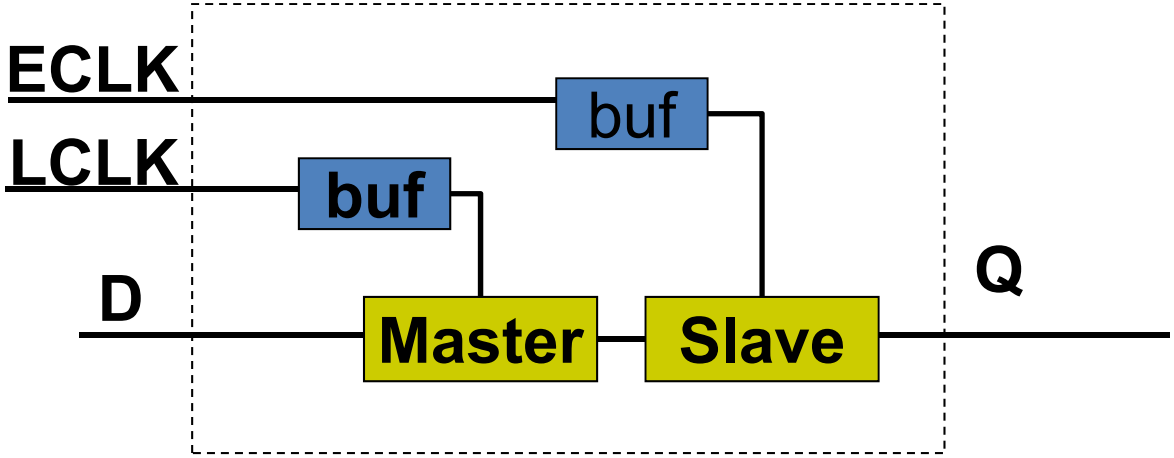
Vt Distribution



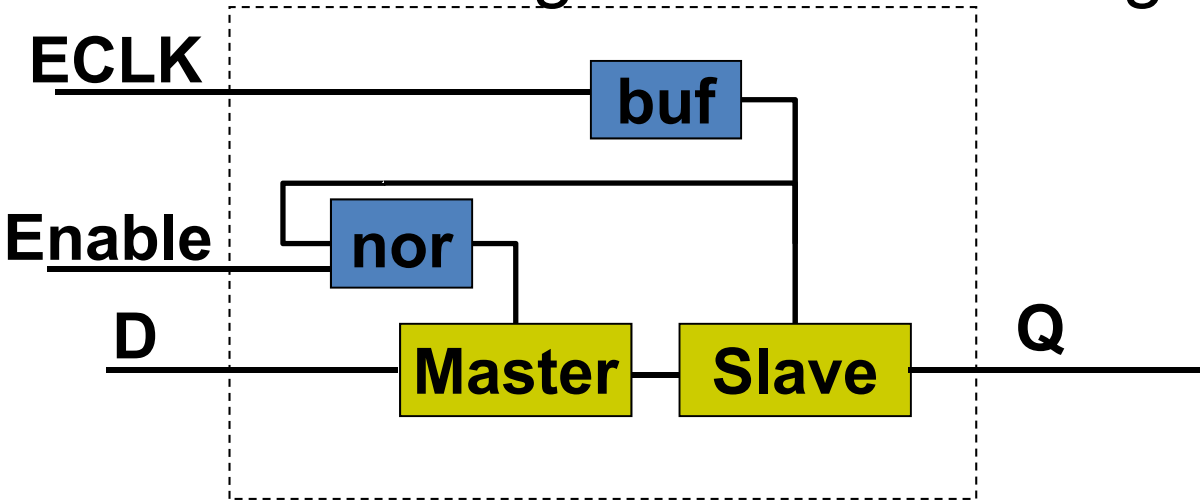
- 4nm channel length bias
- Mid-level Vt workhorse

Single Clock Soft-Edge Flop

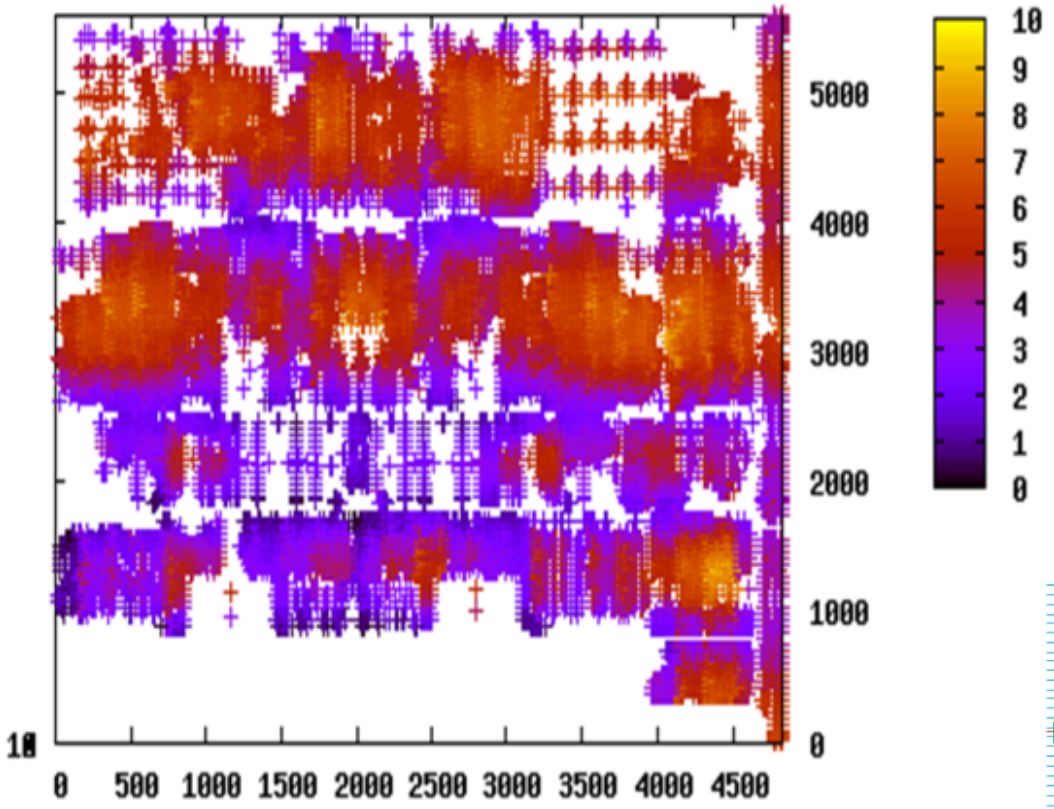
- BullDozer design introduced a dual clock flop



- SteamRoller design moved to a single clock flop

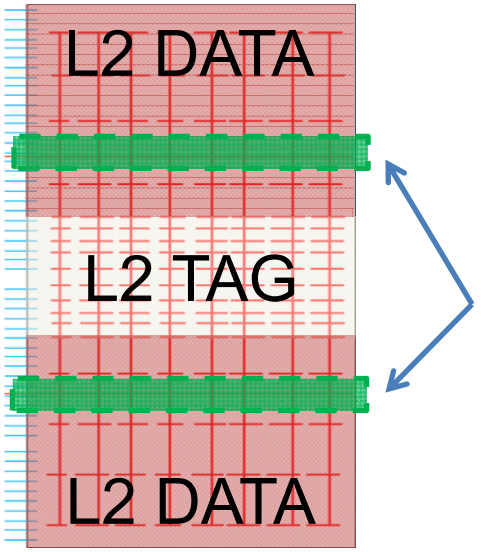


Improved Resonant Clocking



- Added L2 resonant clocks
- Tuned clock gaters

- 6 inductor designs in the core
- 2 inductor designs in the L2 cache



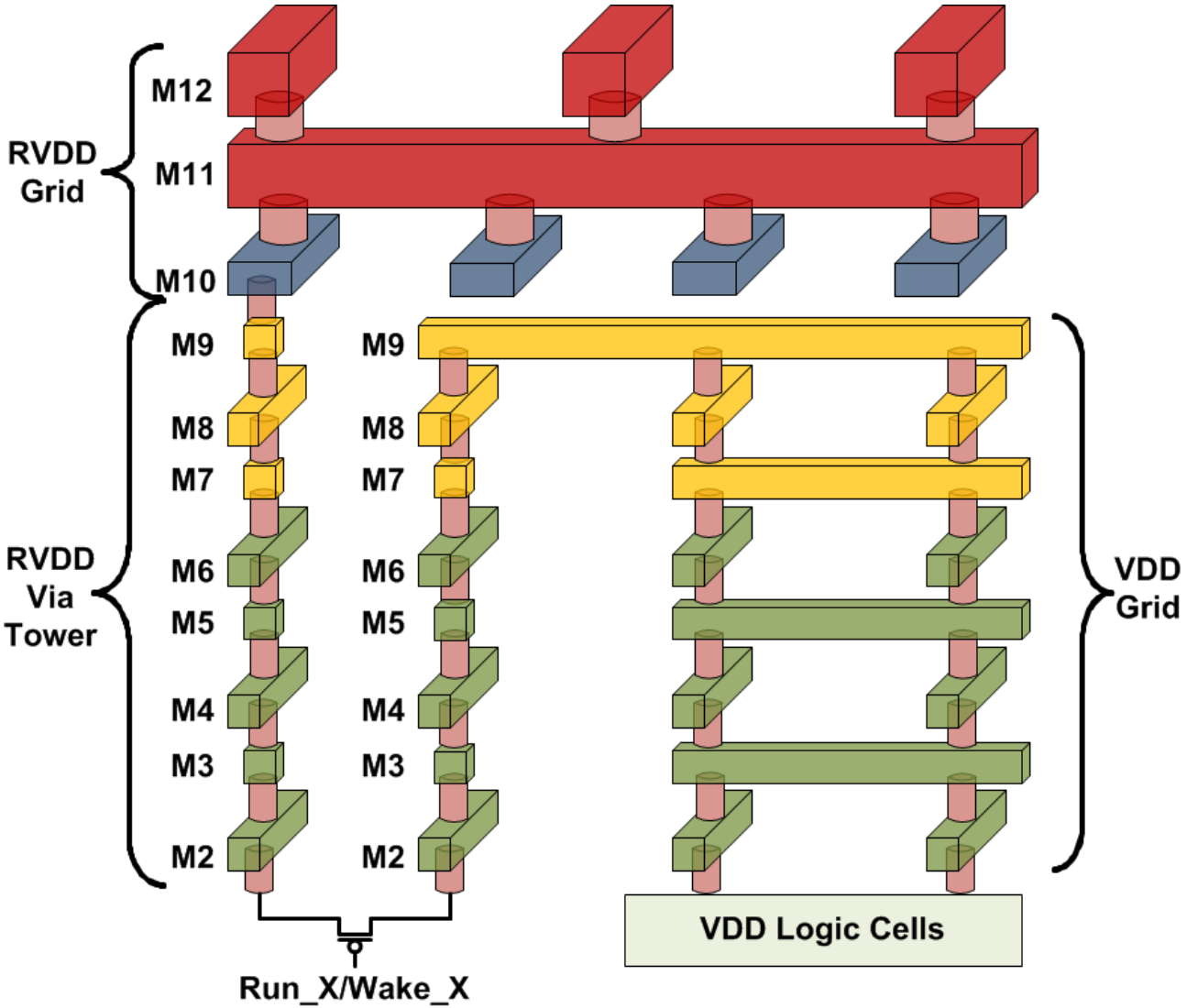
L2 Clock Macros

Distributed Power Gating

- Replaces prior ring implementation
  - Enables L2 way gating
  - Enables future fine-grain power gating
- PMOS headers
- Eliminates power delivery bottleneck
- 90% leakage reduction at high voltage
- Accounts for ~6% core area

Power Gating Grid

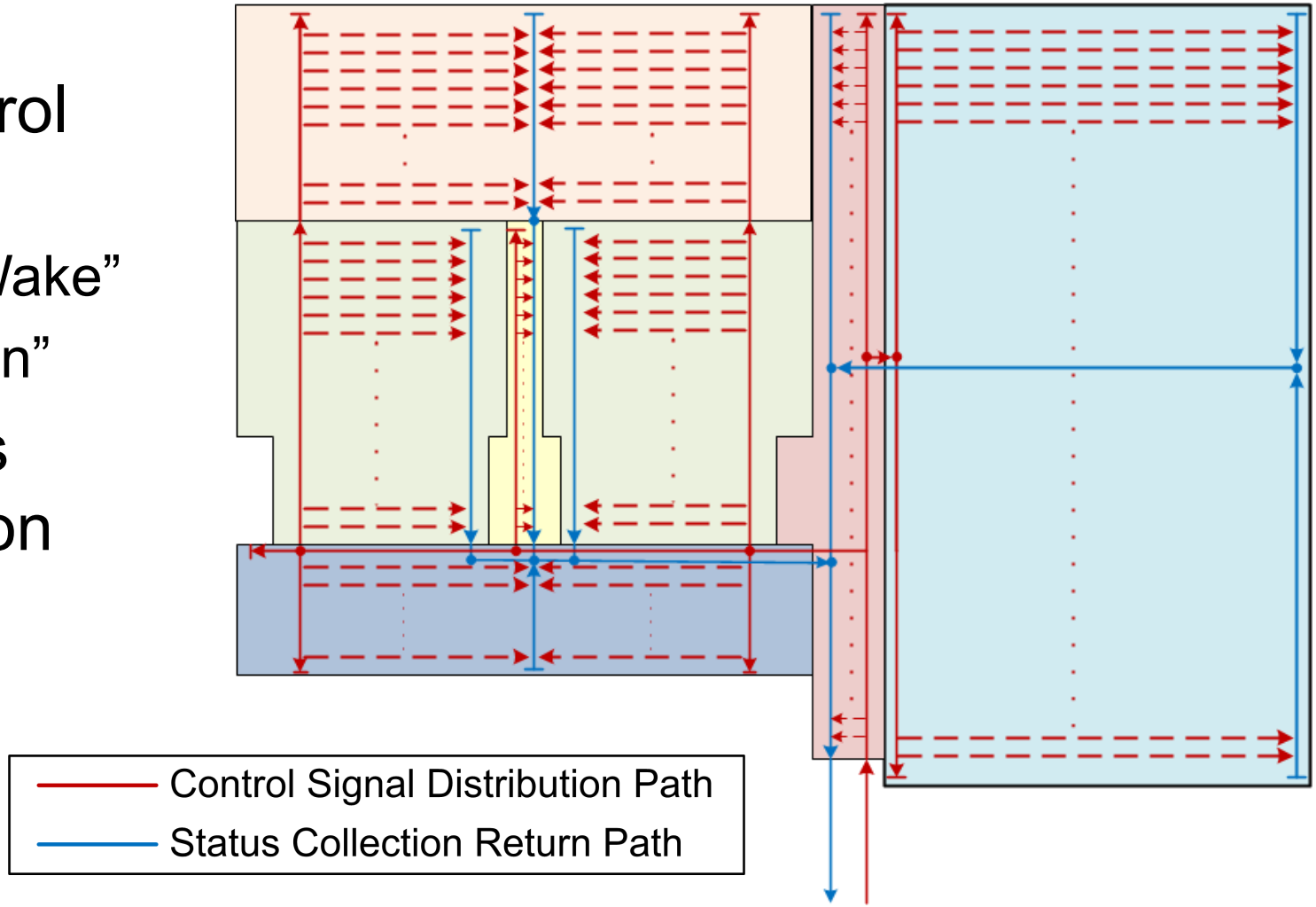
- Always-On supply gridded on thick upper metals
- Gated supply locally distributed on thinner lower metals



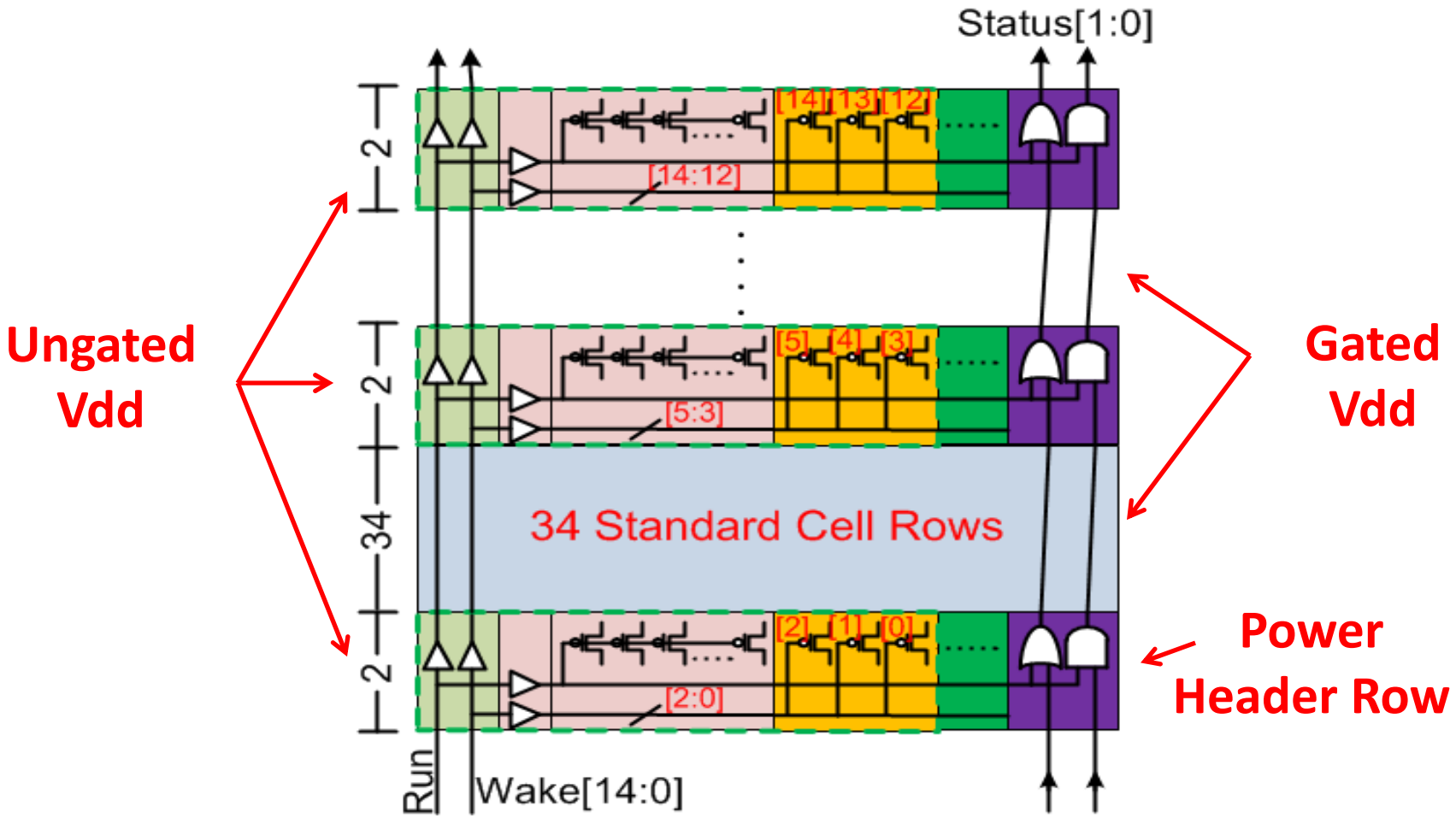
Legend	
Metal Thickness	
	16X
	4X
	2X
	1X

Power Gating Control Distribution

- 16 control inputs
  - 15 “Wake”
  - 1 “Run”
- 2 status collection outputs

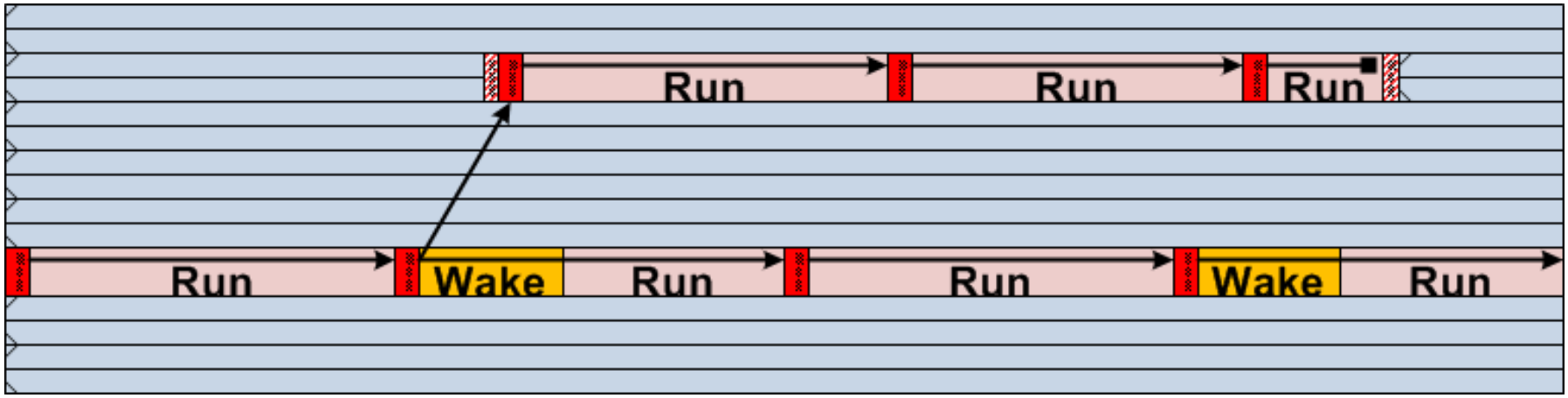


Power Header Construction



- Power gating rows stepped @ pitch of 36 standard cell rows

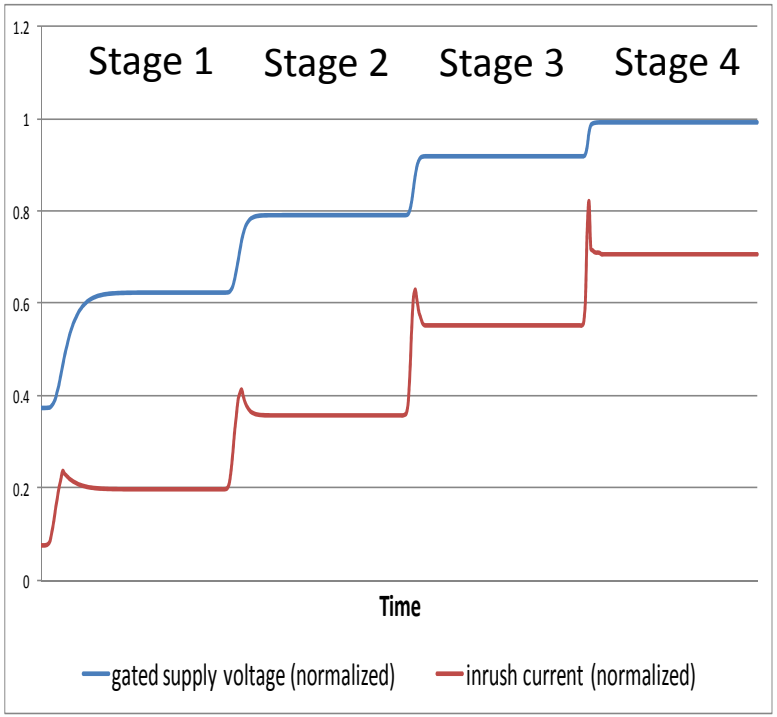
Optional Header Rows



- Additional header rows inserted in areas of high supply droop
  - No placement restrictions
  - No status collection in these areas

In-Rush

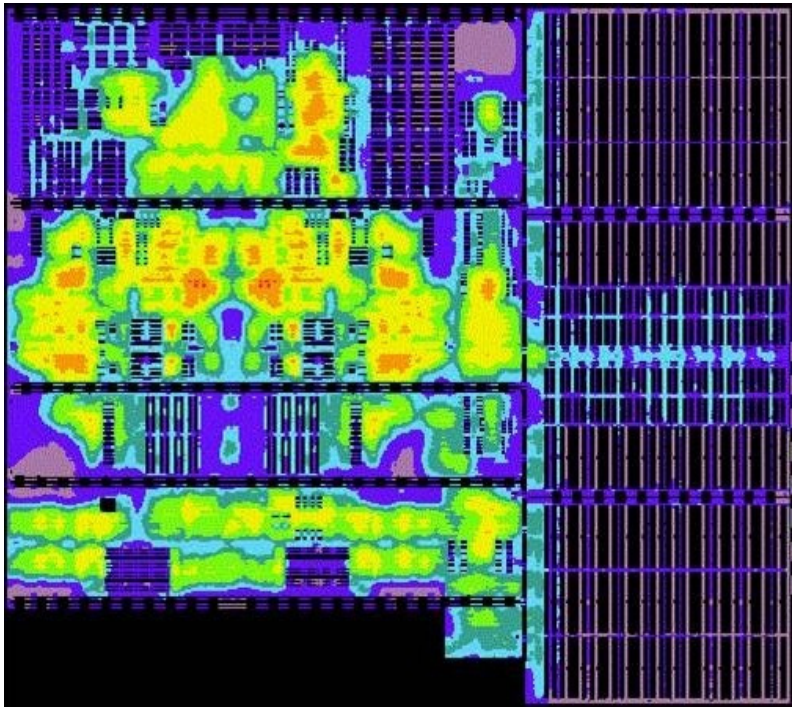
- 4 Wake stages
- Configurable strength and duration for each stage
- 15 strength options



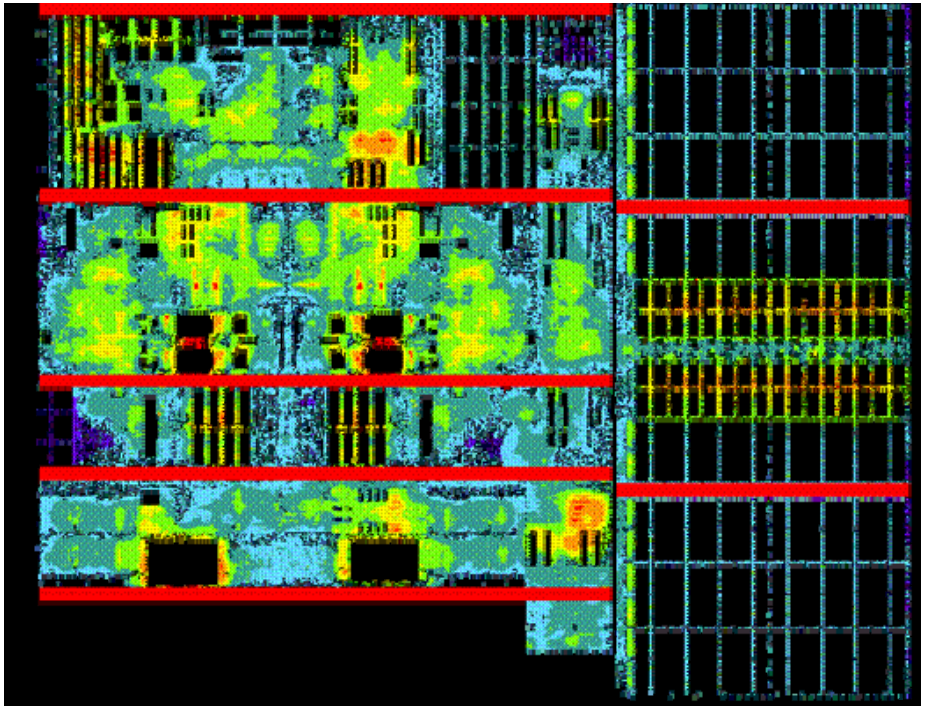
Wake Bit	% Header Enabled	Wake Bit	% Header Enabled	Wake Bit	% Header Enabled
1	0.035	6	0.35	11	2.11
2	0.069	7	0.52	12	3.12
3	0.102	8	0.69	13	4.14
4	0.135	9	0.86	14	5.16
5	0.166	10	1.02	15	6.10

Voltage Droop

- 4% Static Vdd-Vss Compression
  - Includes < 1.5% drop across header

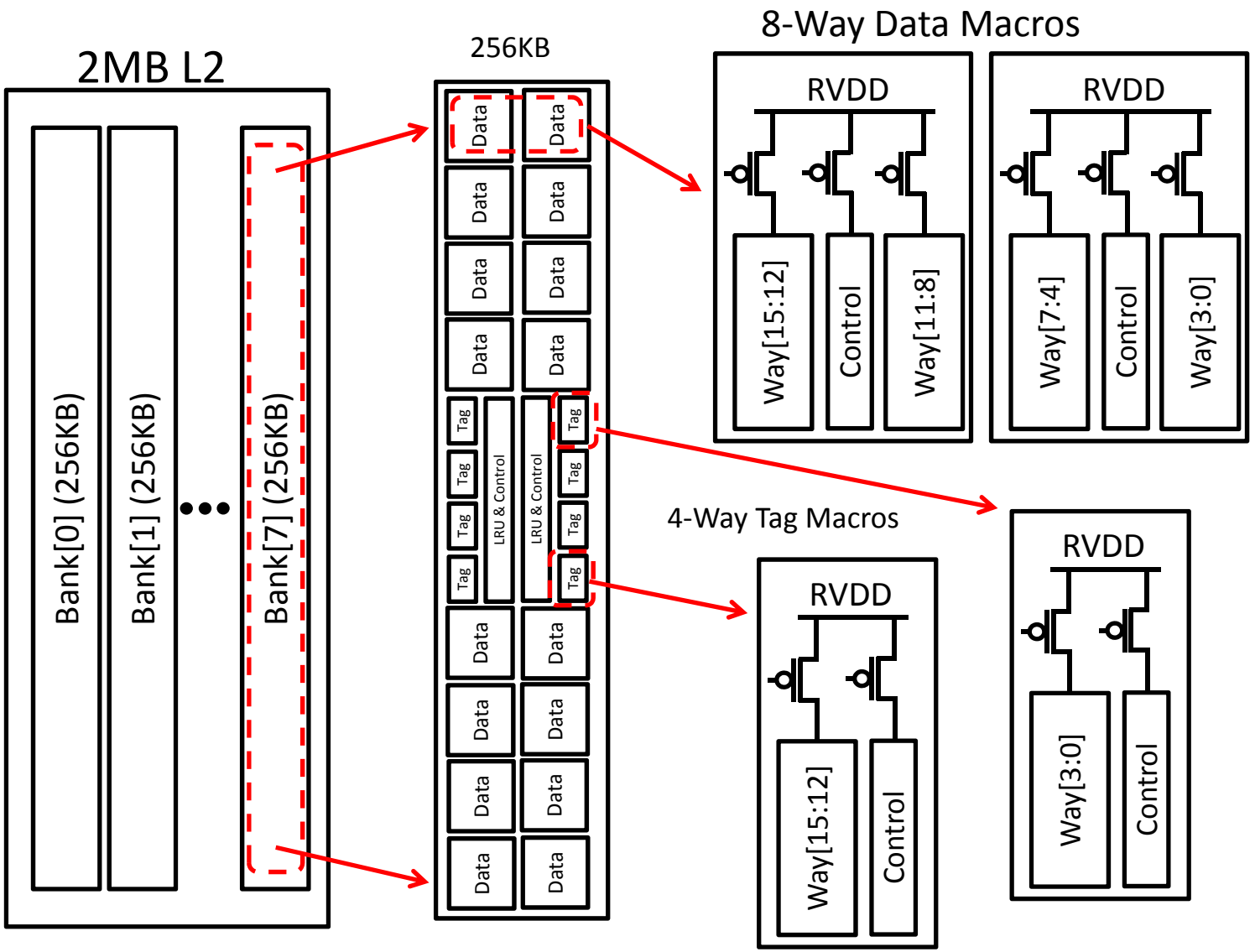


Static Heat map

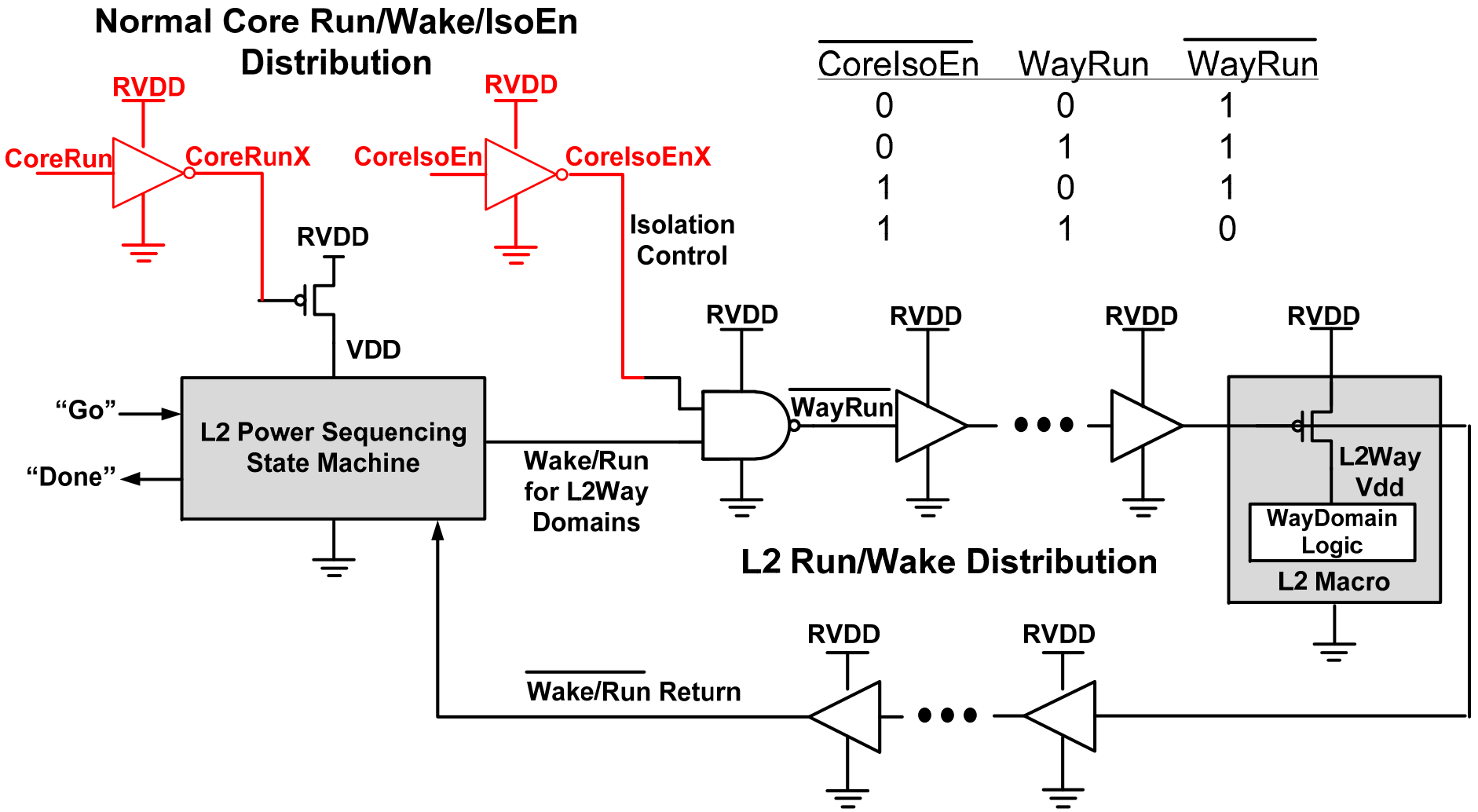


Dynamic Heat map

L2 Way Gating – Power Domains



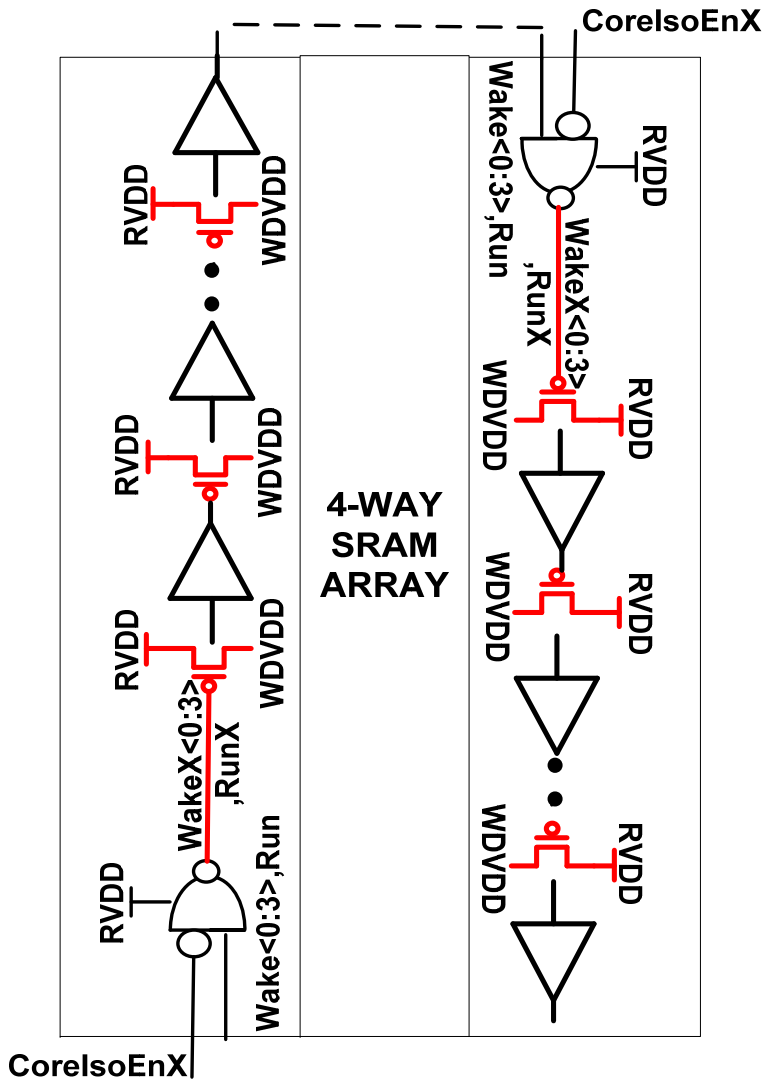
L2 Way Gating – State Machine



CoreIsoEn	WayRun	$\overline{\text{WayRun}}$
0	0	1
0	1	1
1	0	1
1	1	0

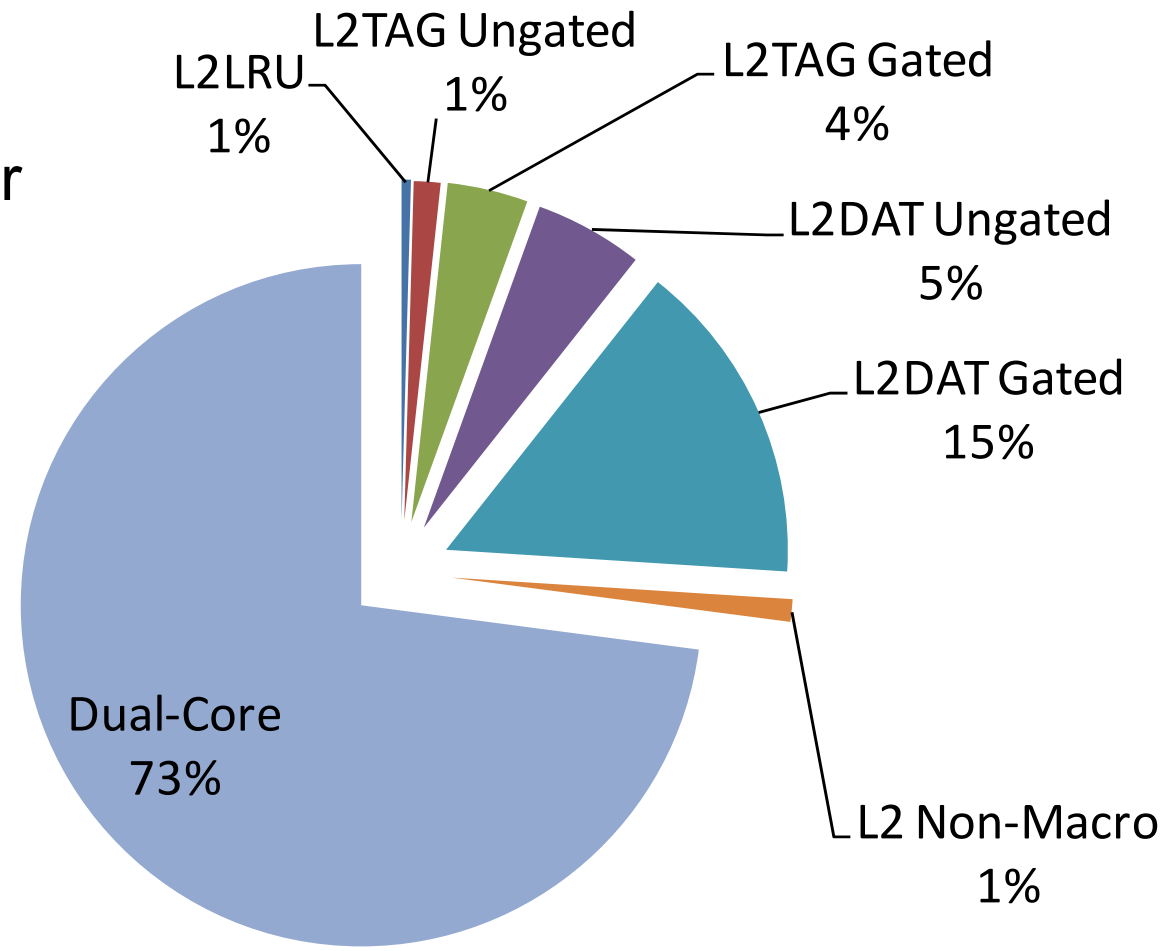
L2 Power Domains & Gating Structures

- Must not allow loss of state retention during gating transition
- 5 Control signals
  - 4 “Wake” = 1x, 4x, 8x, 18x sizes
  - 1 “Run” = 84x size
- Separate L2 state machine
- Way domains divided into two power regions that have independent wake signals.

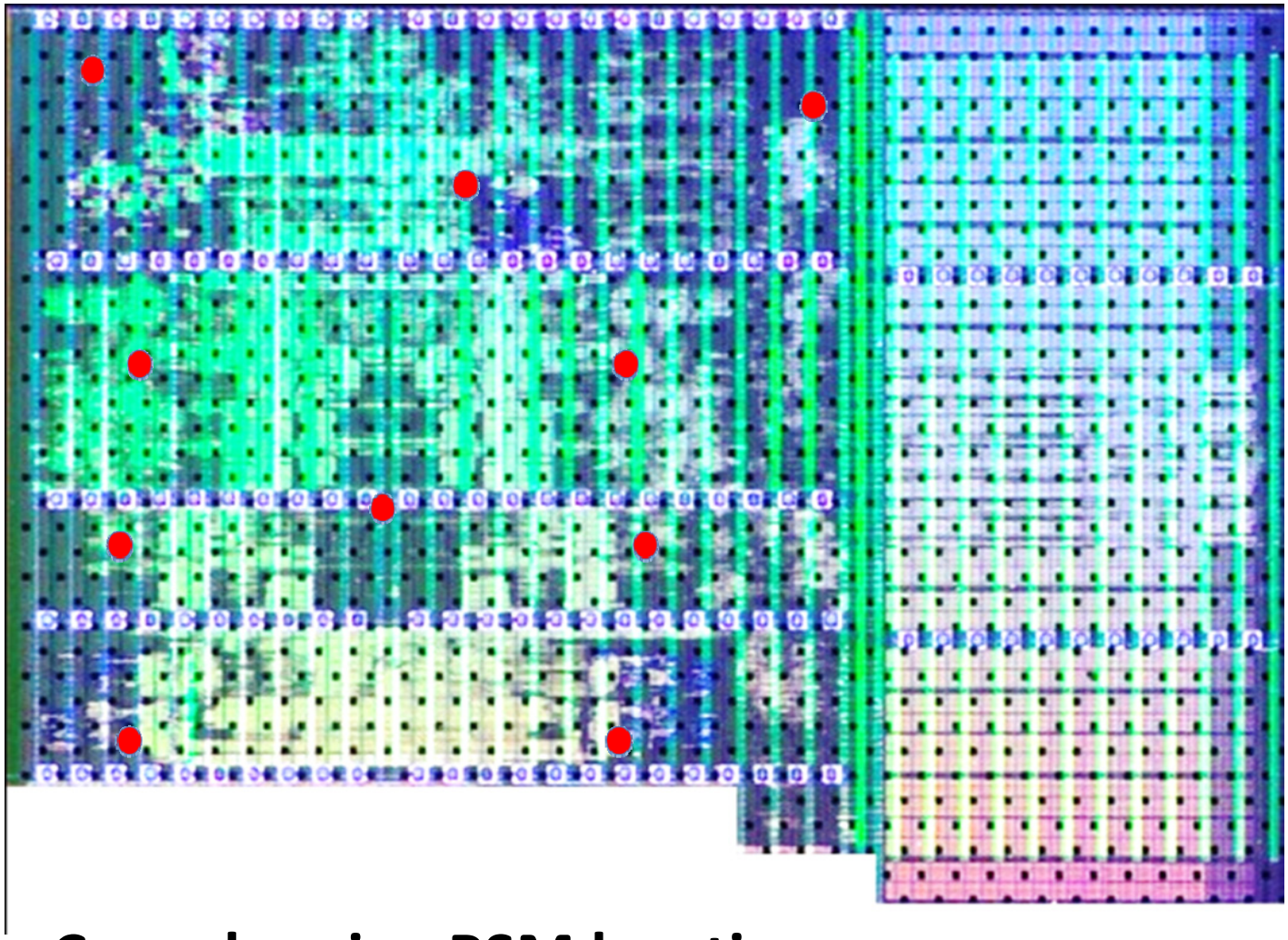


L2 Way Domain Gating Leakage Impact

- L2 Leakage Power
  - 27% of total
  - 5% per WD
- L2 Active Power
  - 6% of total
  - 0.5% per WD

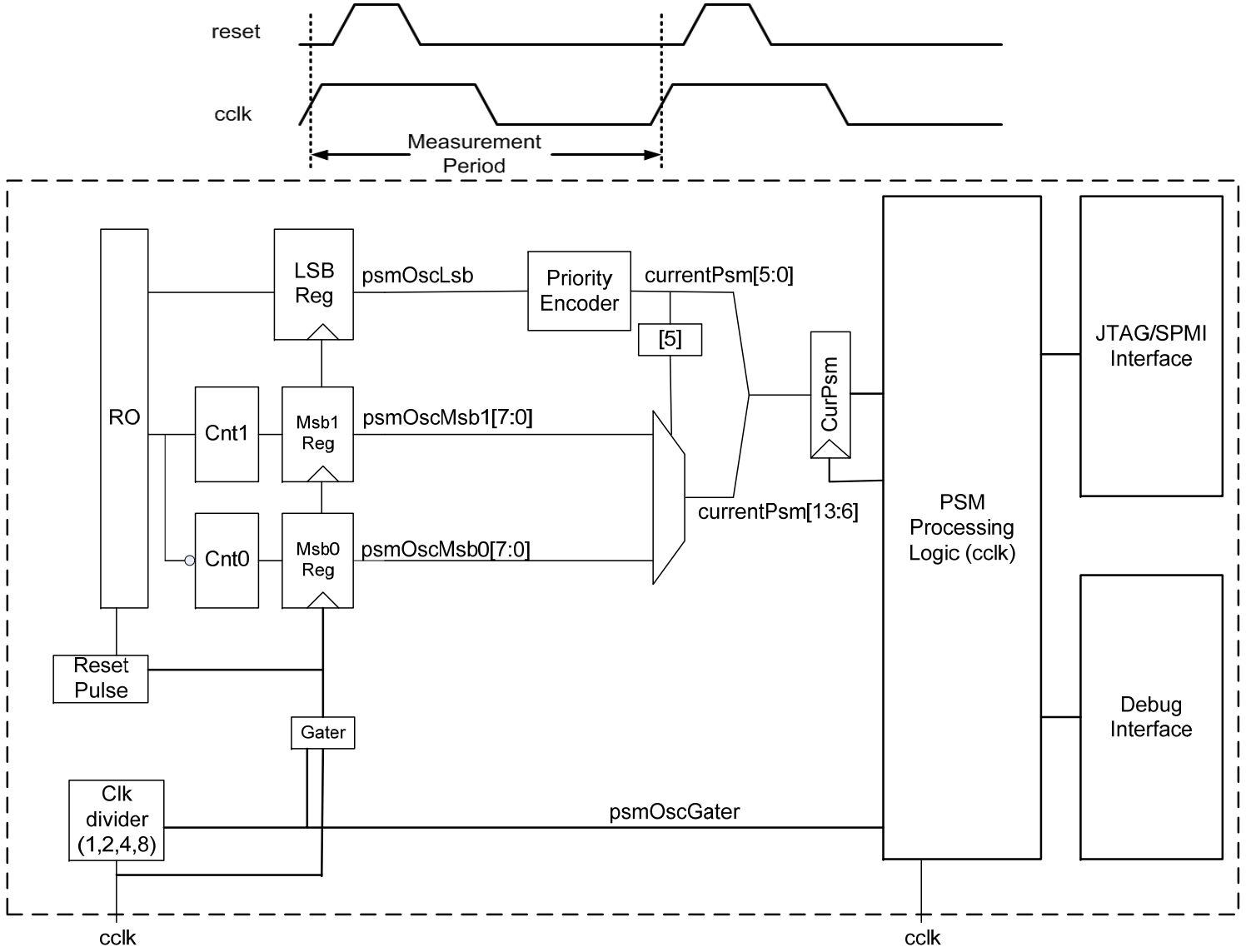


Power Supply Monitor (PSM)

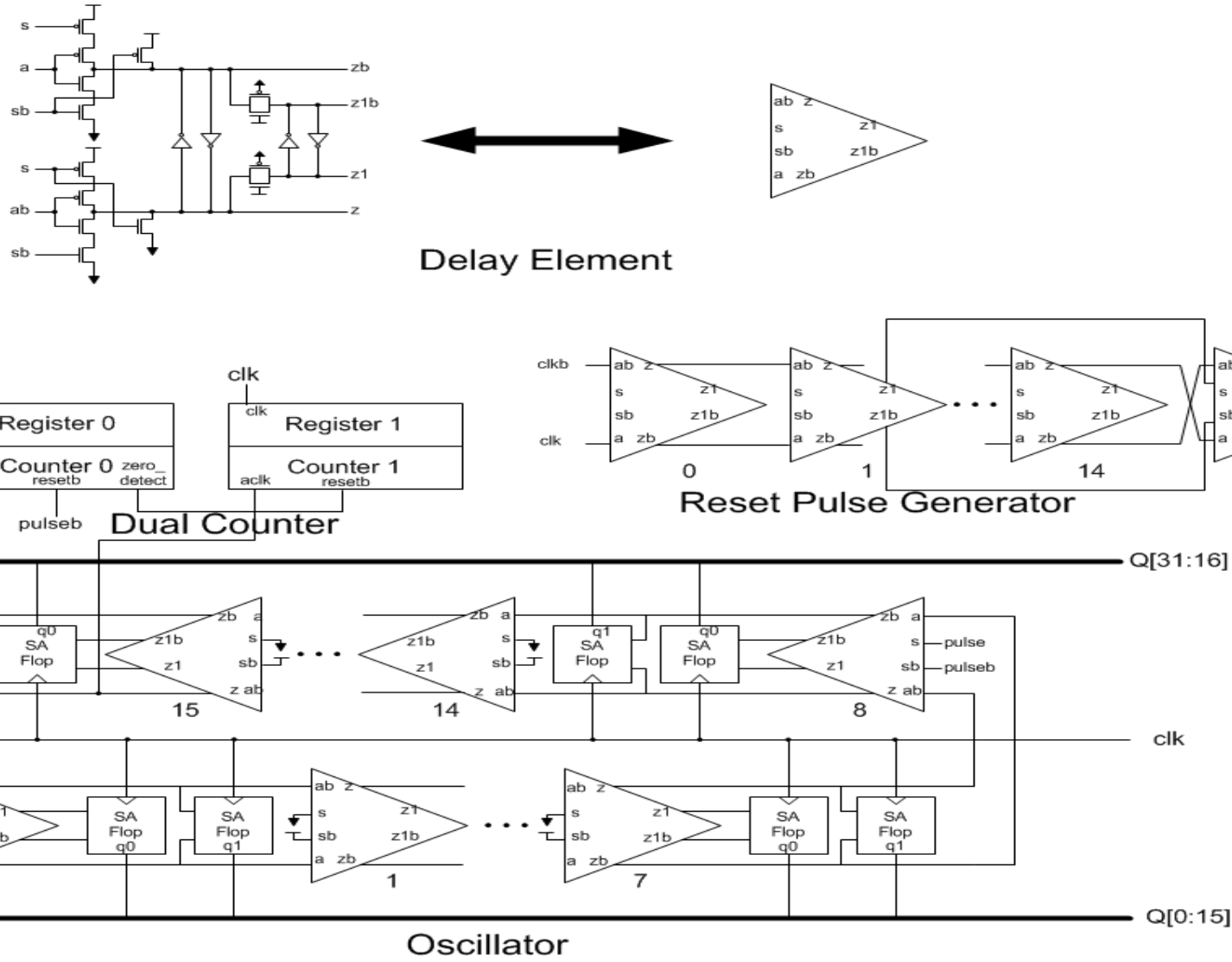


Core showing PSM locations

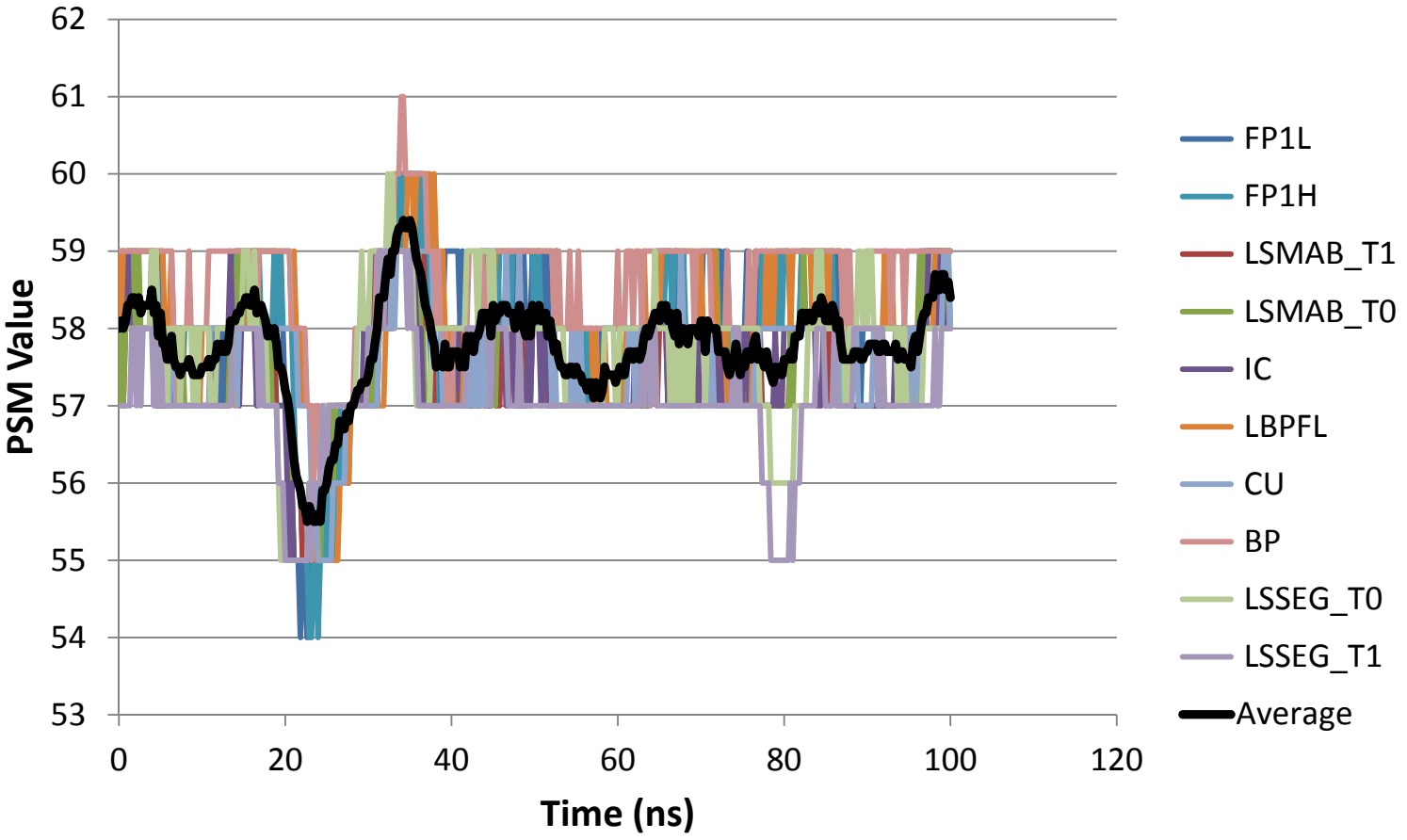
PSM Top level diagram



PSM Components



Single Core Step Response Measurement



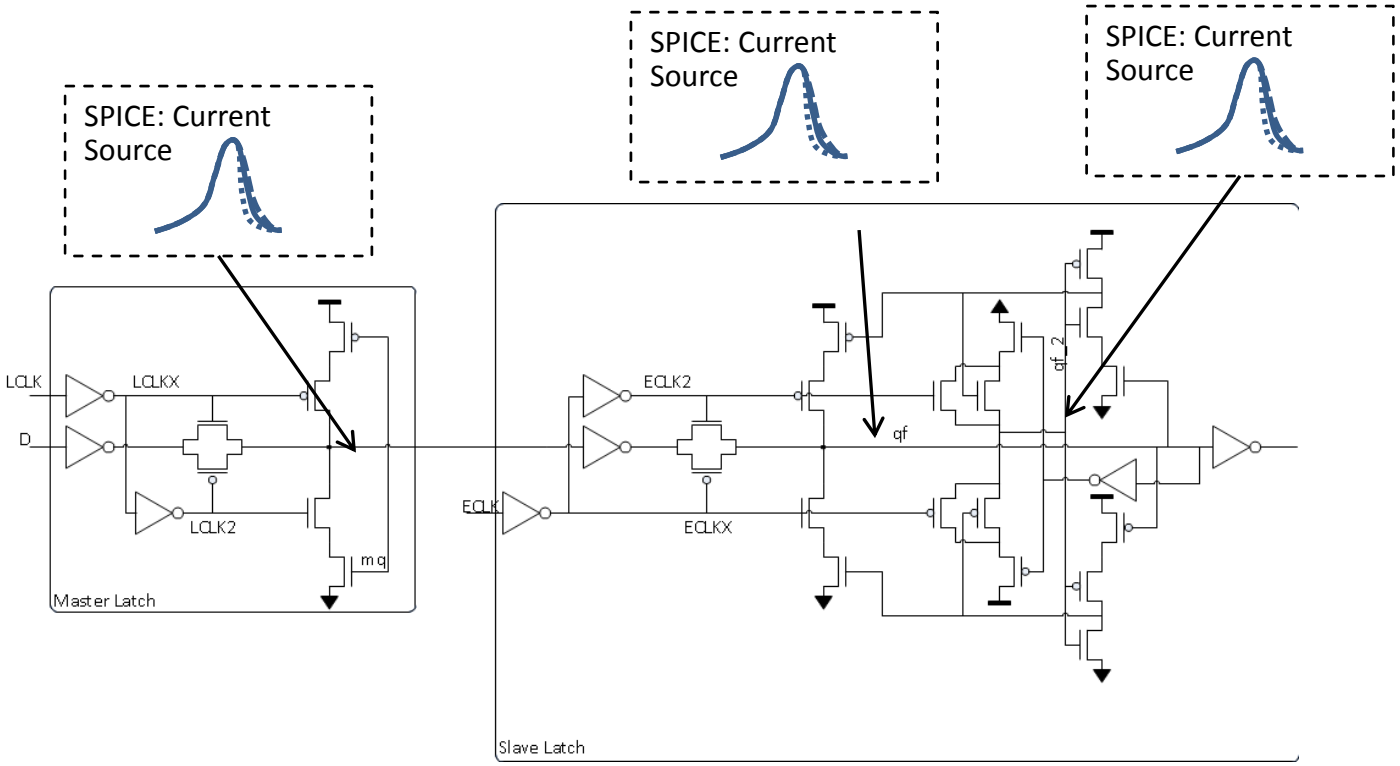
Clock cycle = 3.8GHz, 1.4V, 100ns step-load trace
(Clock Stretcher Off)

Soft Error Flow

- Bulk process more susceptible to Soft Errors than SOI processes used in prior designs.
- Process
  - Calculate SER FIT for each flop in the design
  - Define vulnerability scaling factors
  - Determine SRAM array vulnerability
- Changes:
  - Convert highly vulnerable flops to Error Tolerant flops
  - Add parity protection to some array structures
  - Creation of Error-Tolerant “Lite” flops with improved SER FIT

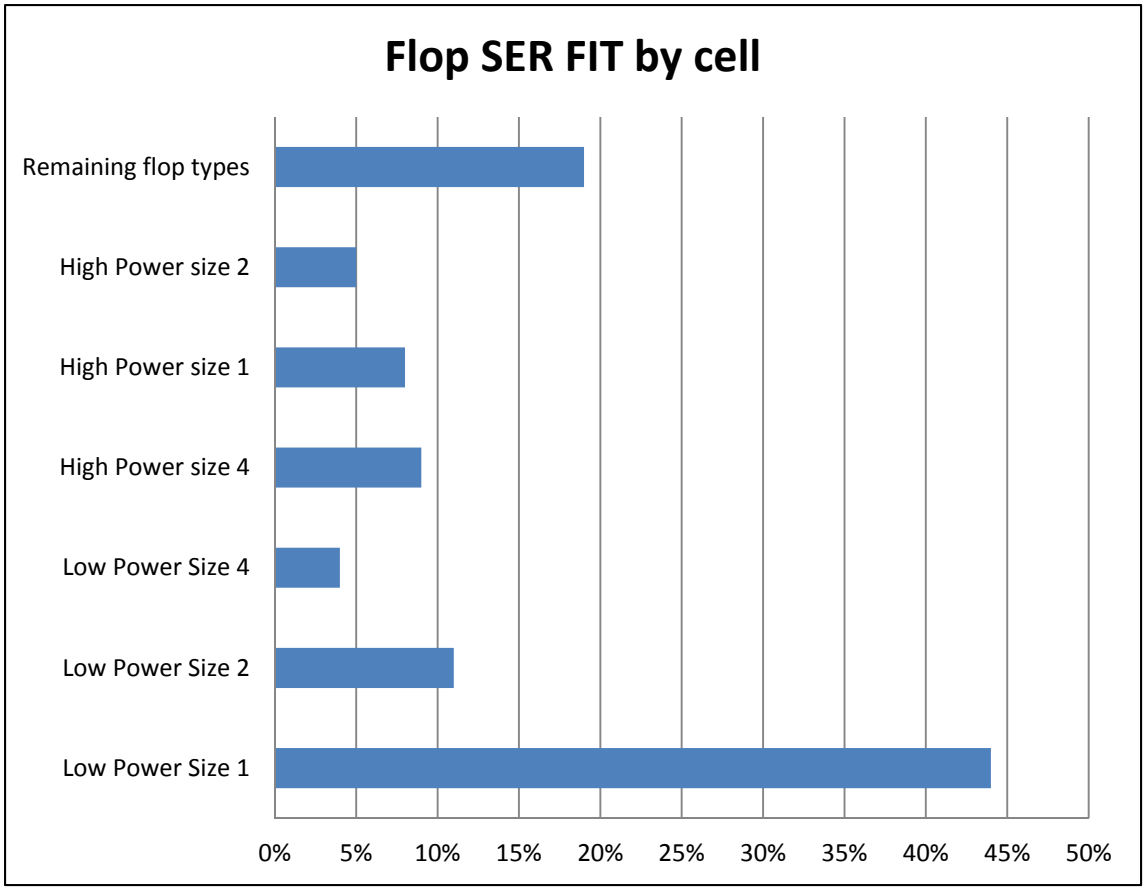
Flop Qcrit Simulations

- Evaluating flop SER FIT was done by determining the critical charge (Qcrit) required to flip a storage node in the flop.



Error Tolerant Flop

Soft Error FIT Contributions



- Created more robust cell variants for most cells

Conclusion

- 28nm High Performance HKMG Bulk core
- Successfully mitigated technology impacts
- Improved designer efficiency
- Substantial power reduction
  - Distributed power headers
  - L2 way domain gating
- Developed power supply monitors
- Mitigated susceptibility to soft errors

Adaptive Clocking System for Improved Power Efficiency in a 28nm x86-64 Microprocessor

Aaron Grenat, Sanjay Pant, Ravinder Rachala, Samuel Naffziger

Outline

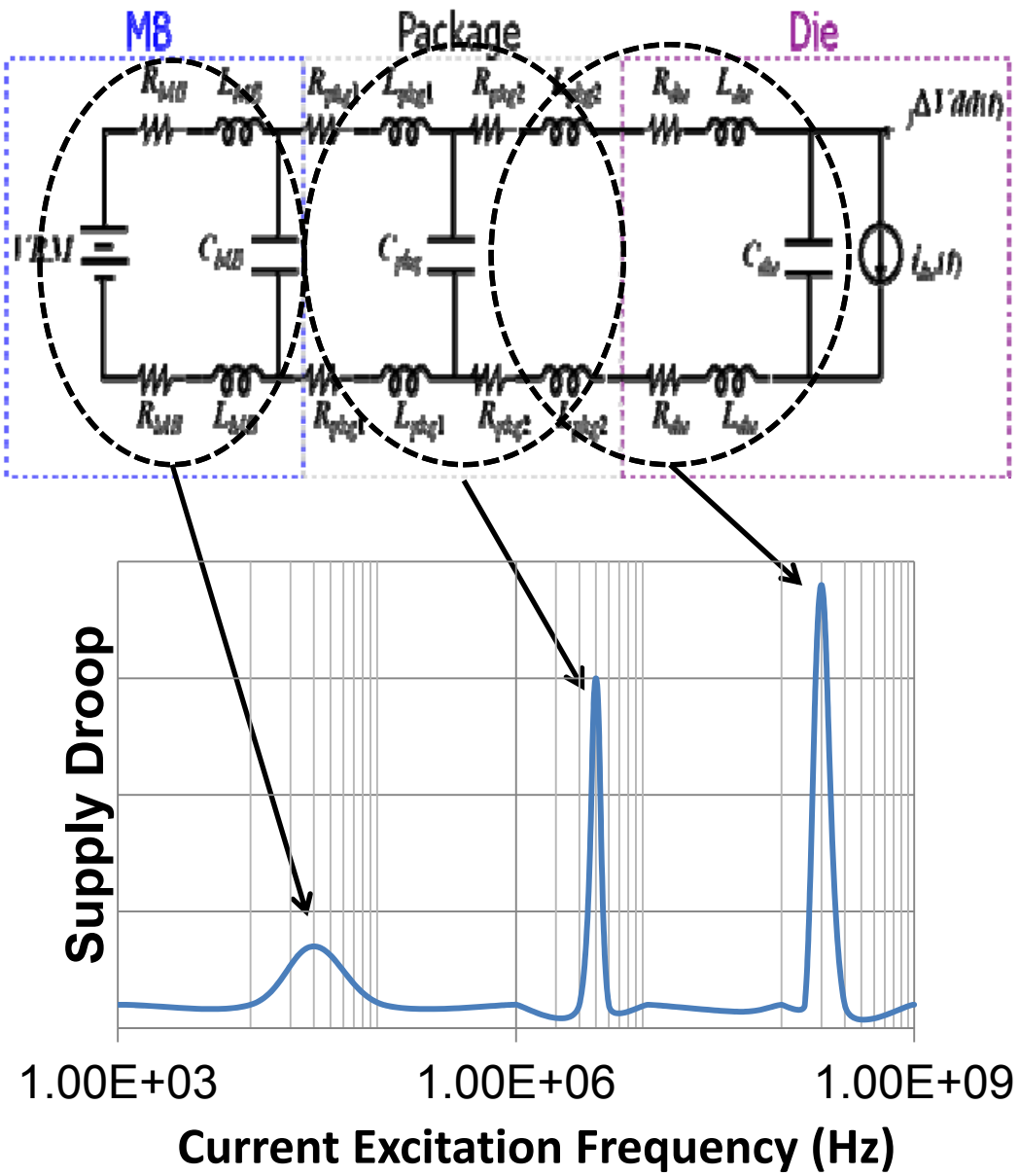
- Impacts of droop on microprocessor core power
- Adaptive clocking overview
- Circuit Operation
- Measured results
- Conclusion

Cause of Voltage Droop

- Voltage droop is caused by a change in current though the package inductance
- Excitation frequency of the droop is defined by:

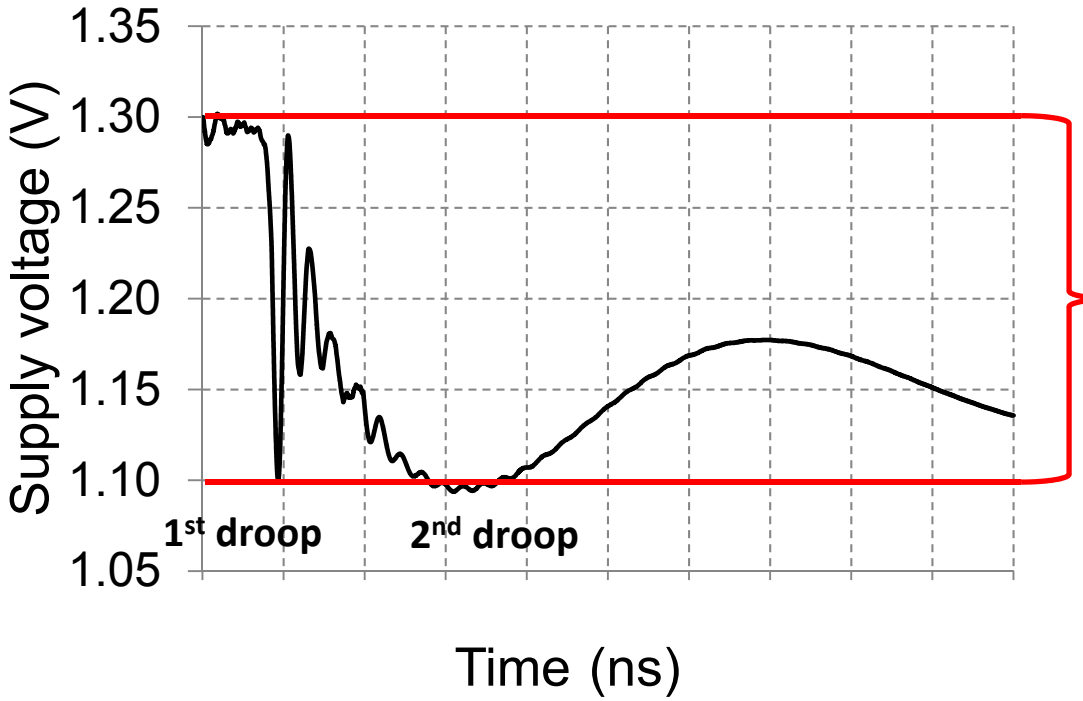
$$f = \frac{1}{2\pi\sqrt{LC}}$$

- Droop Interaction:
  - 1st -> Die and Package L
  - 2nd -> Package
  - 3rd -> Mother Board



Impact of Droop

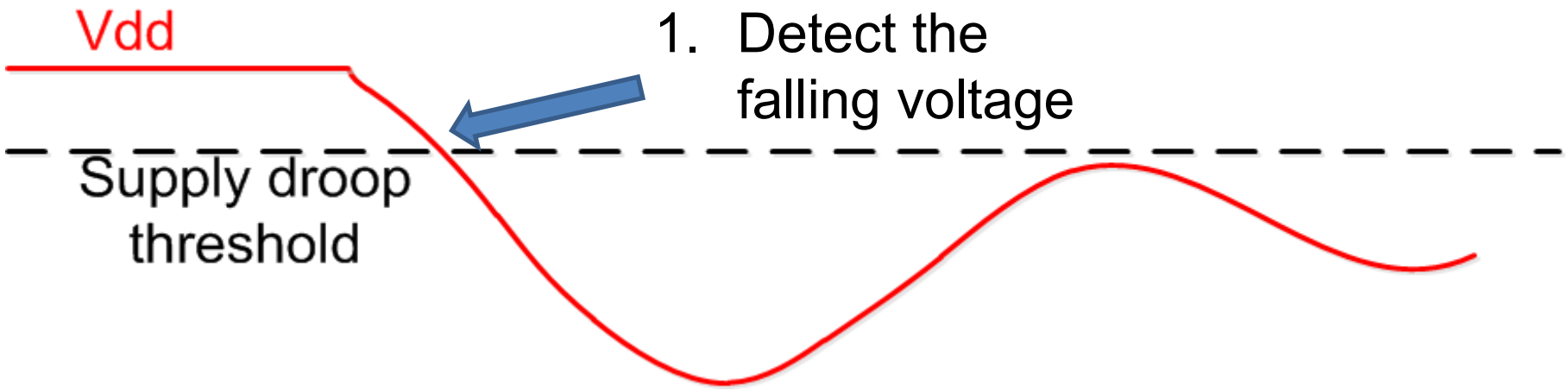
- Below: A current step is applied resulting in 200mV of di/dt*L induced voltage droop
- Microprocessor maximum frequency is determined by the voltage seen at the lowest point in the droop



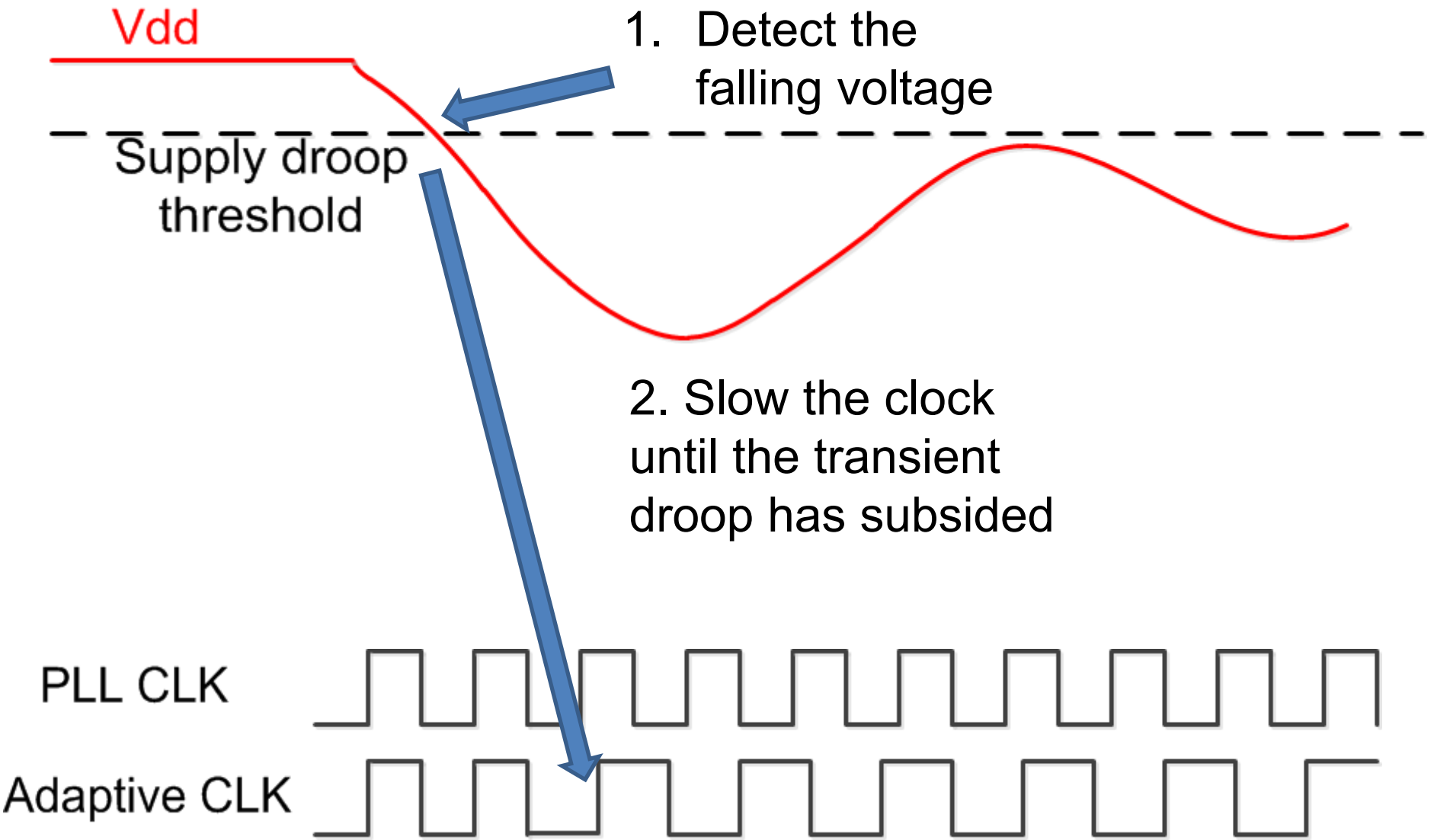
Voltage droop has a significant impact to power and frequency.

Reducing the droop impact increases performance/watt!

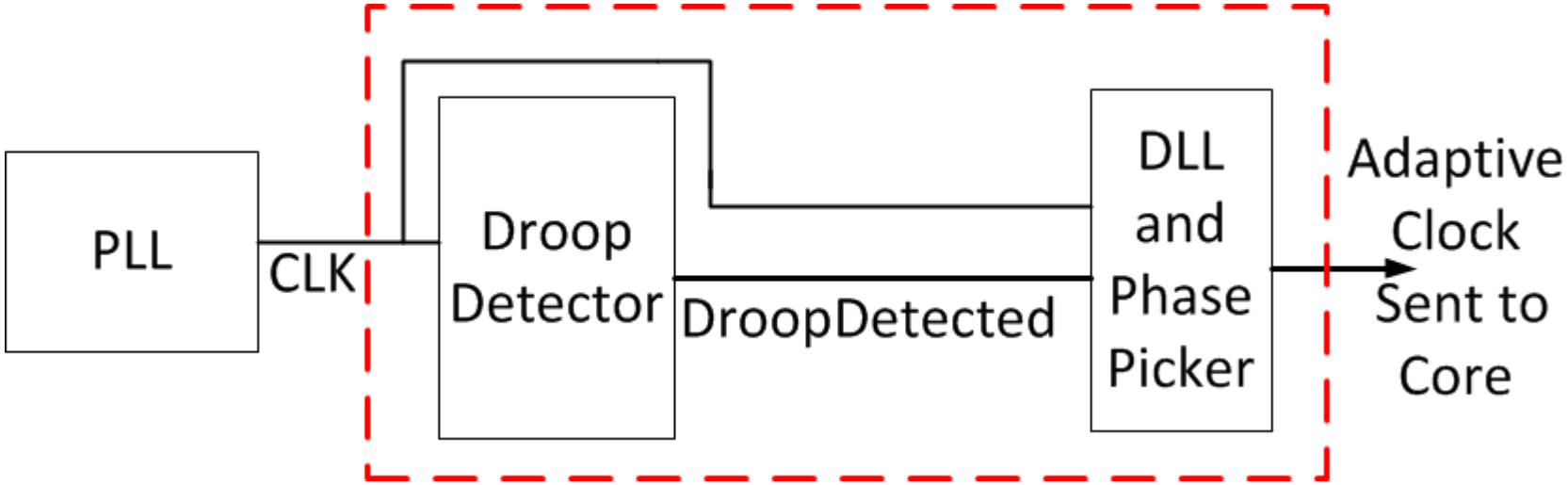
Adaptive Clock Overview – Step 1

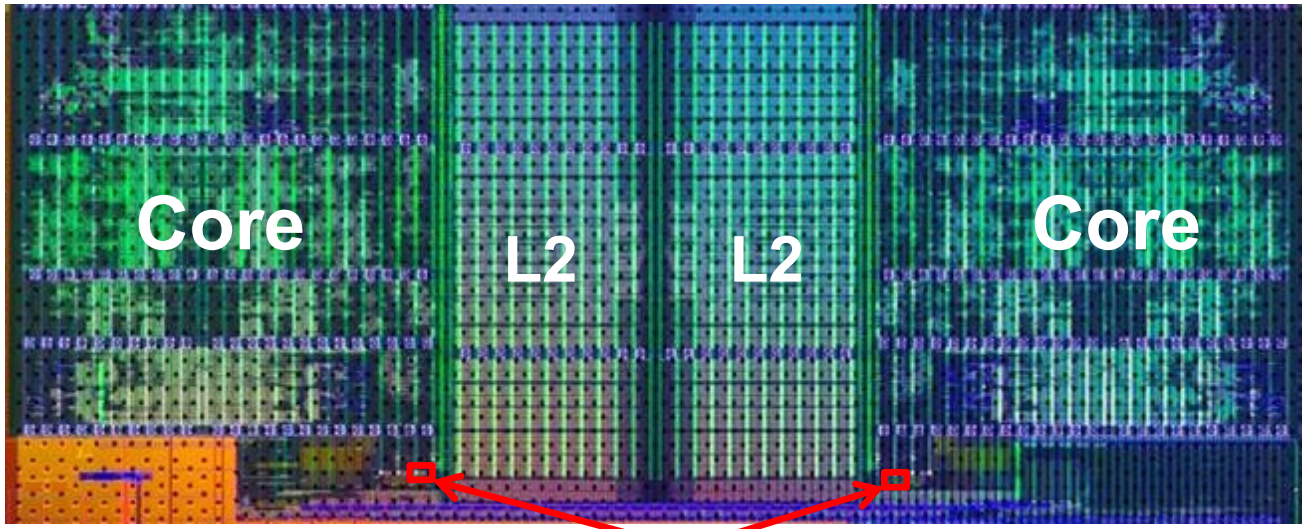


Adaptive Clock Overview – Step 2

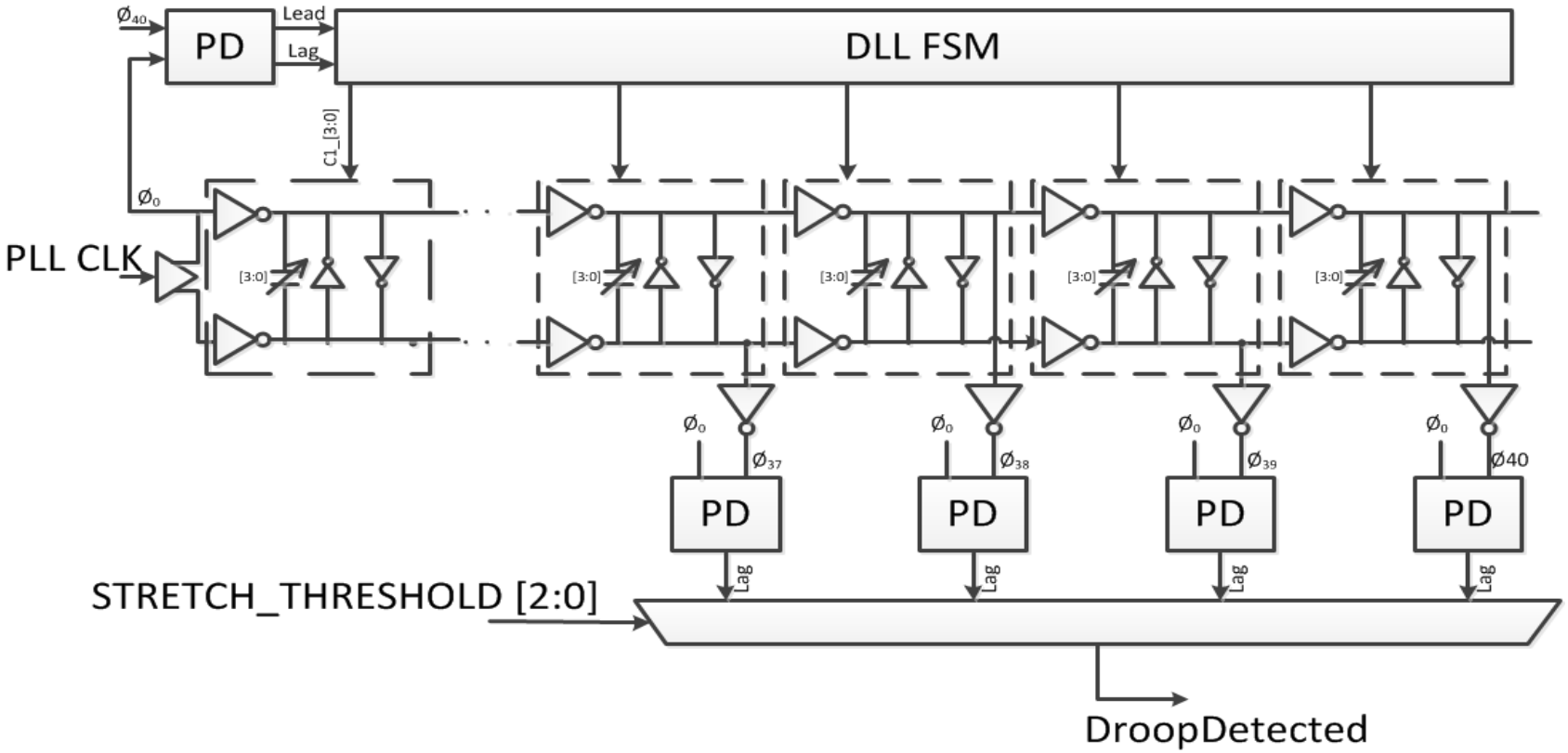


Adaptive Clock System



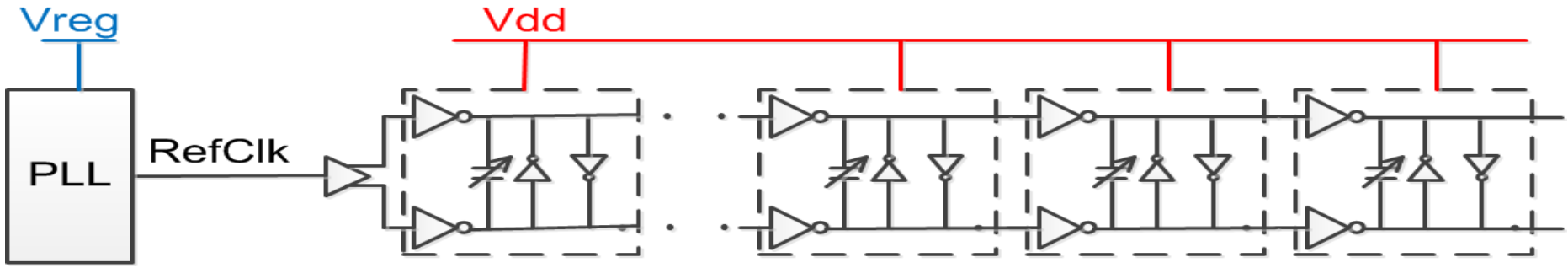


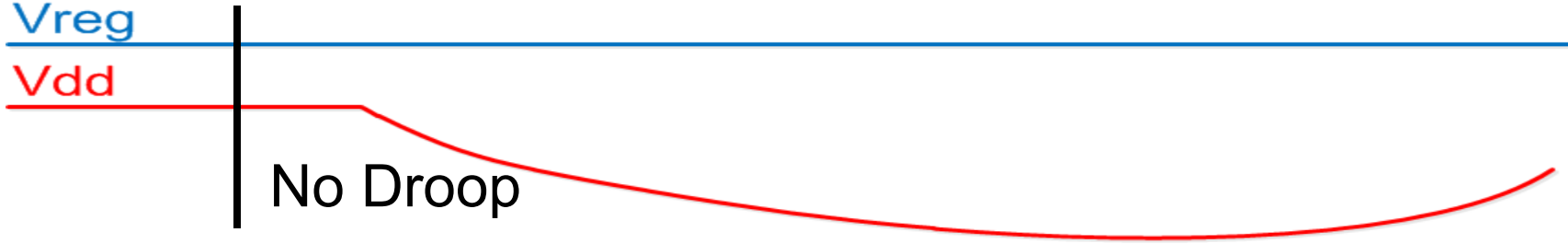
Droop Detector Overview

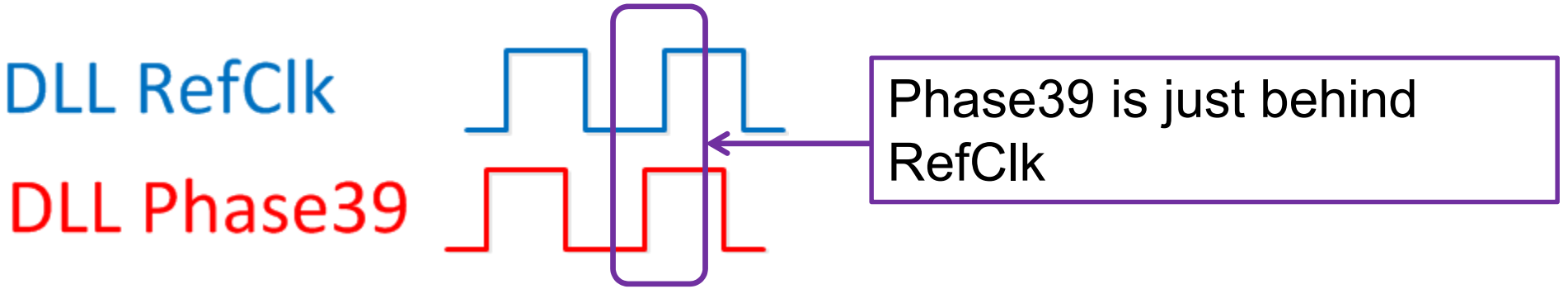


1) A DLL generates clock phases on the noisy supply
2) As the voltage droops, the phases will shift w.r.t. the reference clock which is generated on a clean supply
3) A phase detector (PD) detects that the DLL phase is lagging the reference and triggers DroopDetected

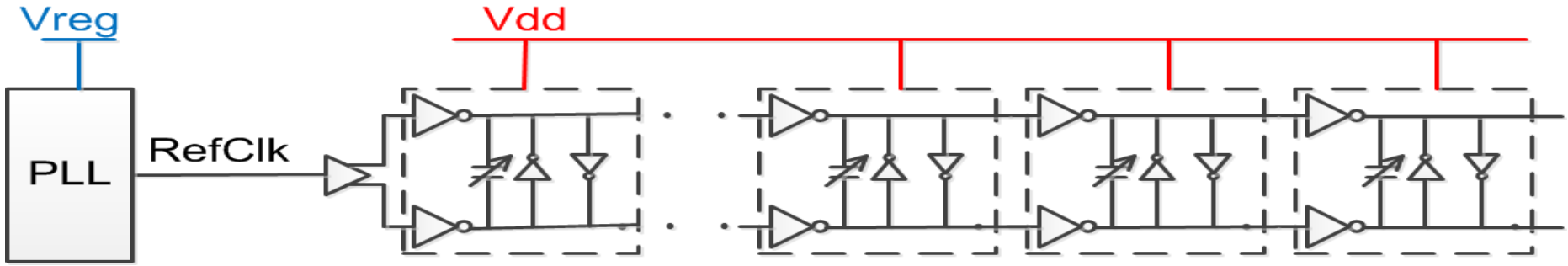
Droop Detector - No Droop

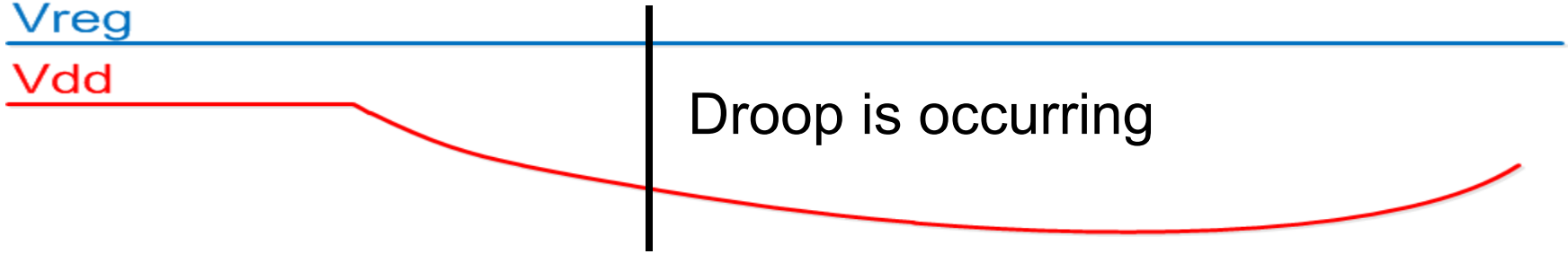






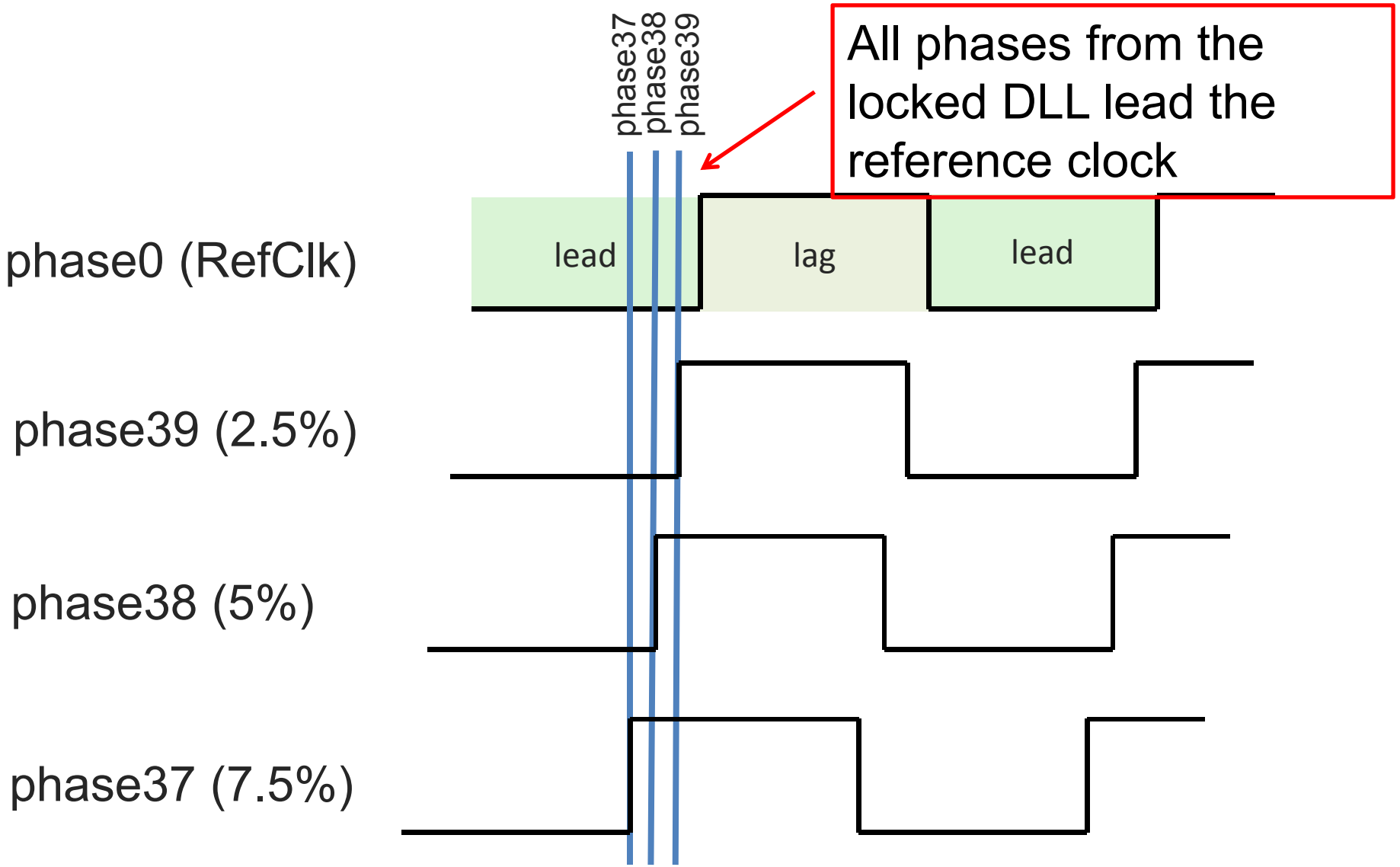
Droop Detector - Droop Occurs



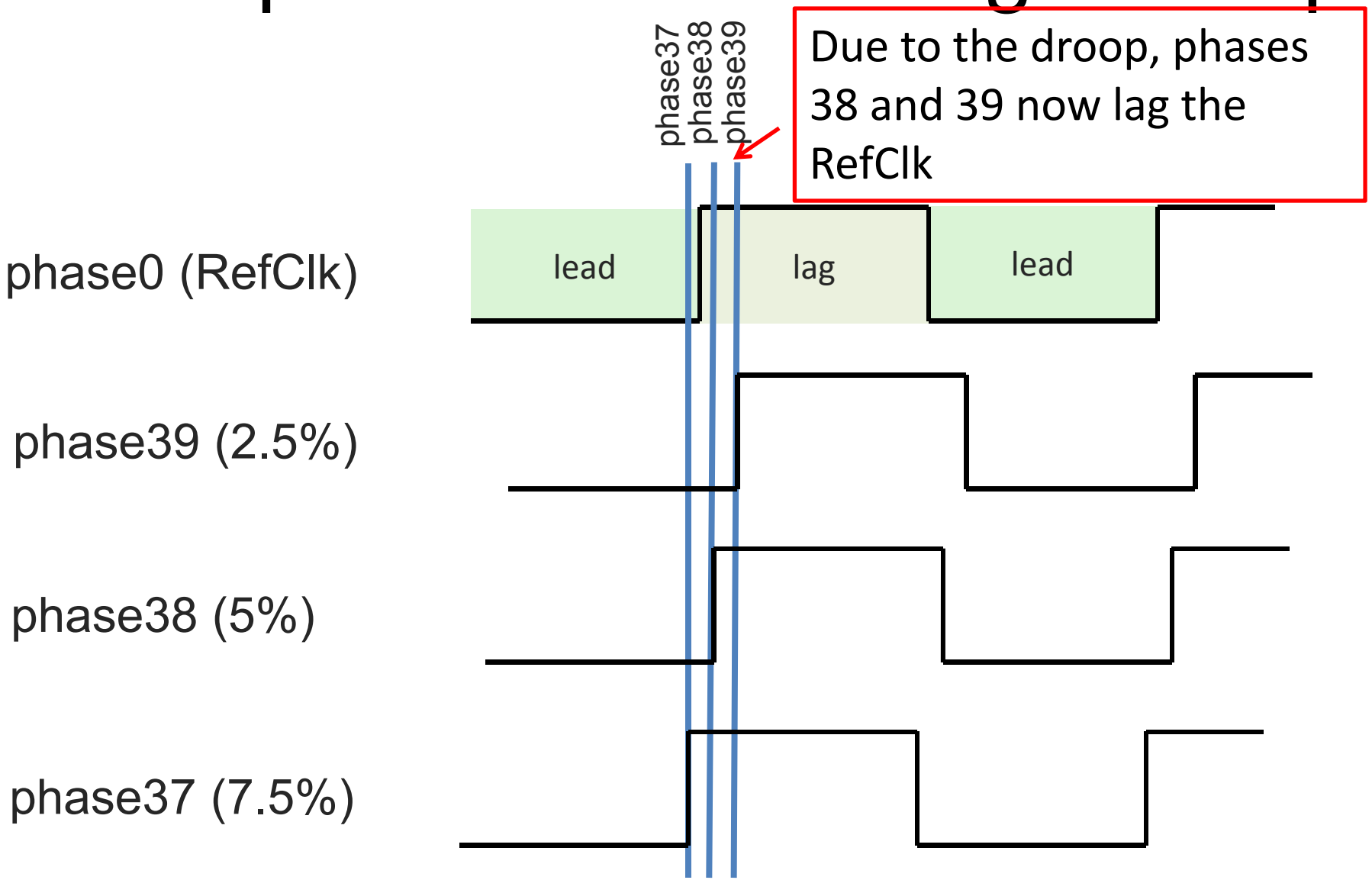


DLL phase39 leads RefClk due to the decreased voltage on the DLL and resulting increase delay per stage

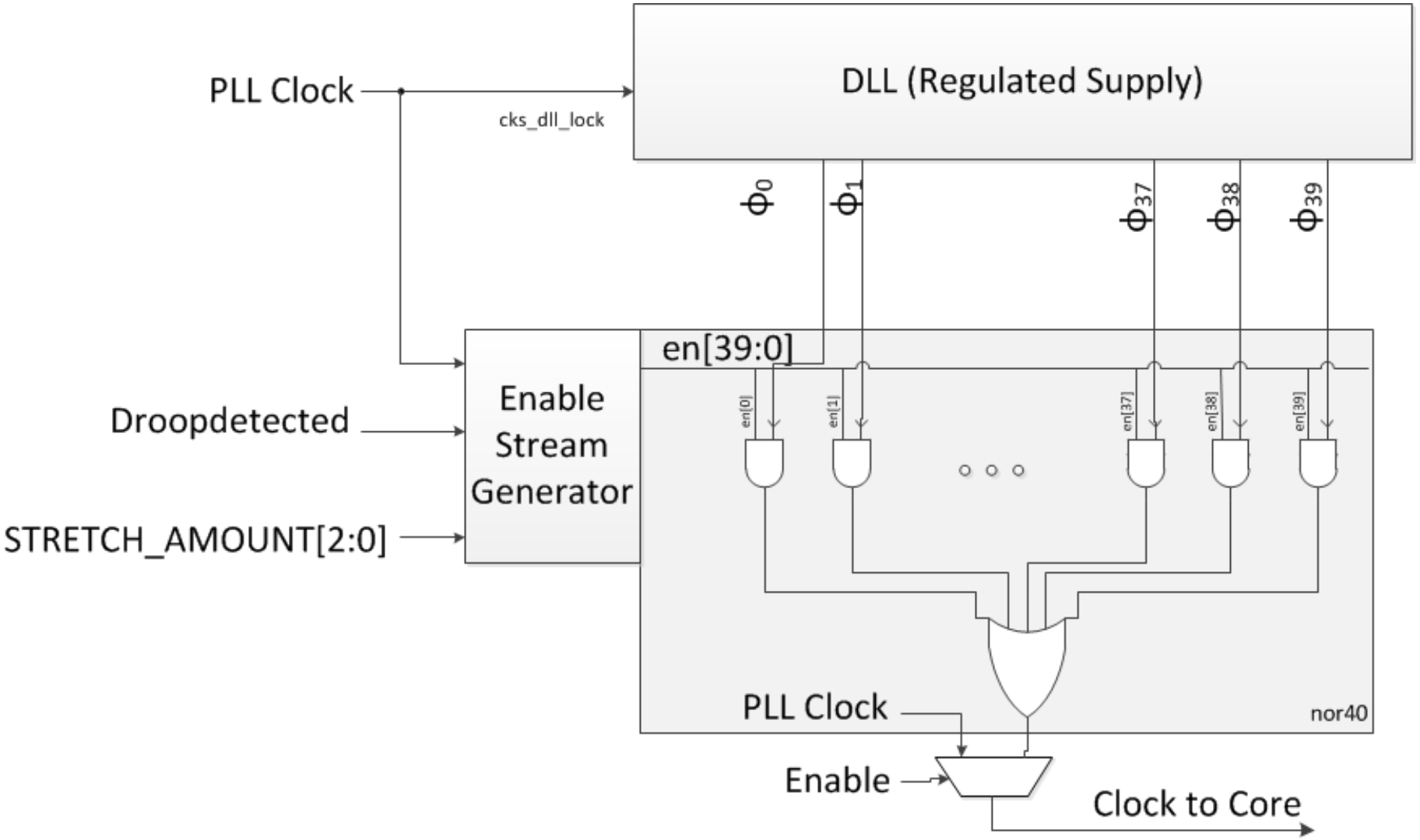
Droop Detection – No Droop



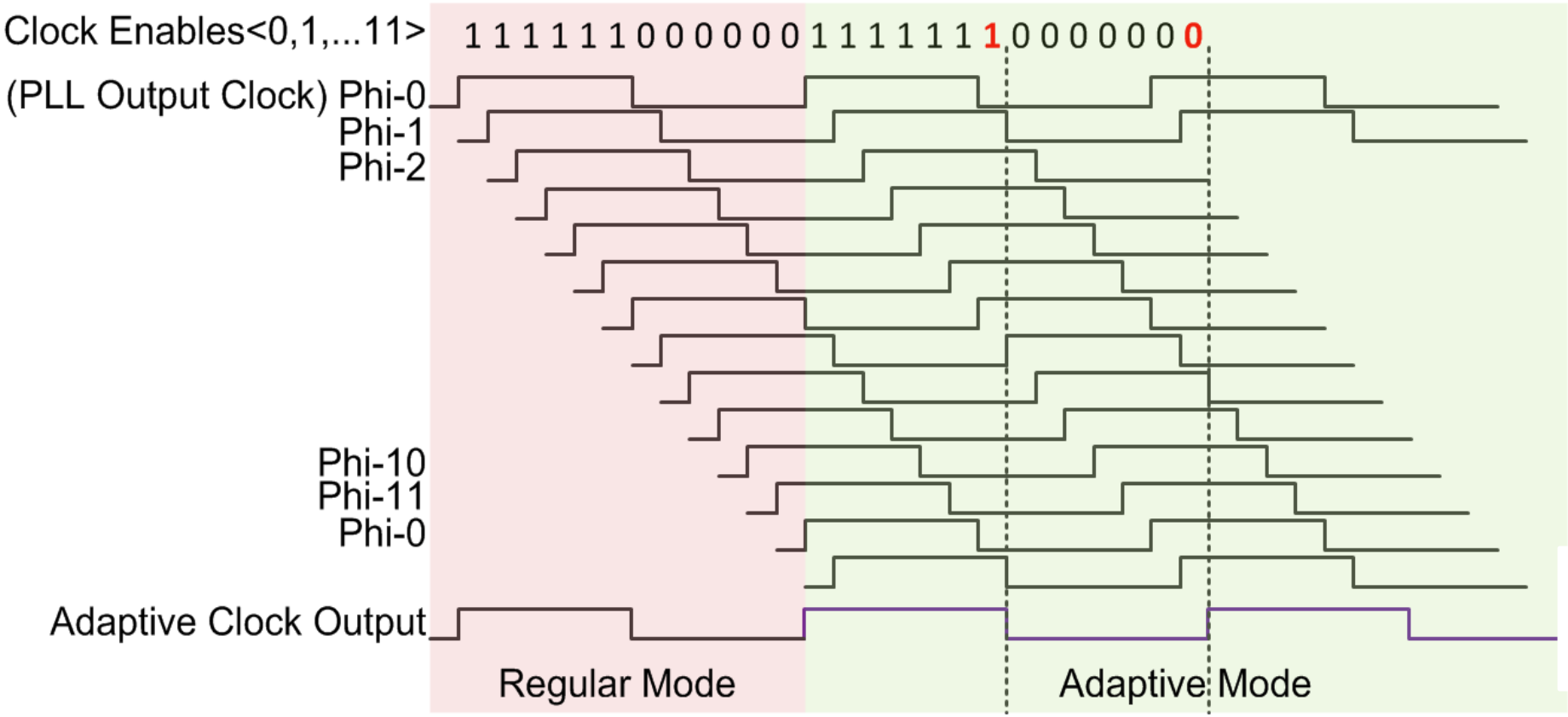
Droop Detection –Voltage Droop



Adaptive Clock Generator

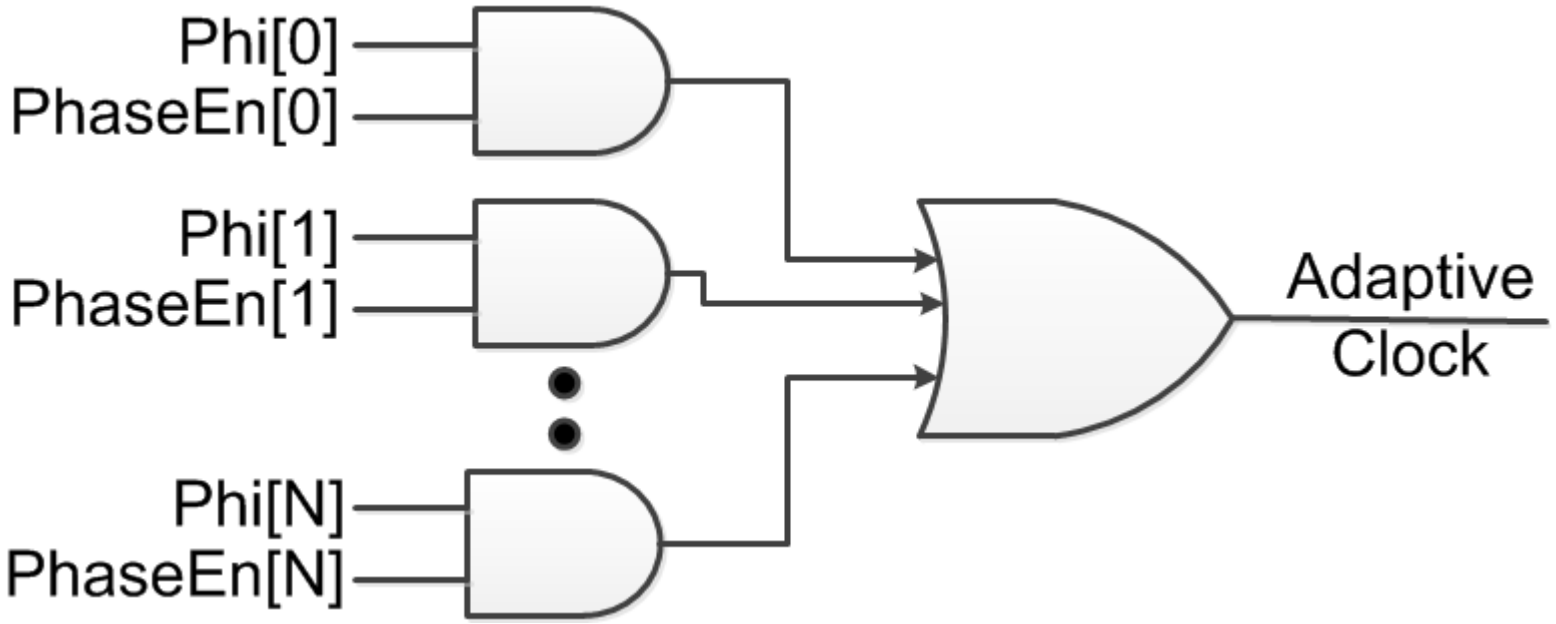


Adaptive Clock Phase Picking



- A DLL Creates 40 phases (12 shown above)
- The Clock Enables “pick” the phases to create the clock
- 40 phases provides 5% frequency resolution for slowing the clock

Adaptive Clock Phase Picker



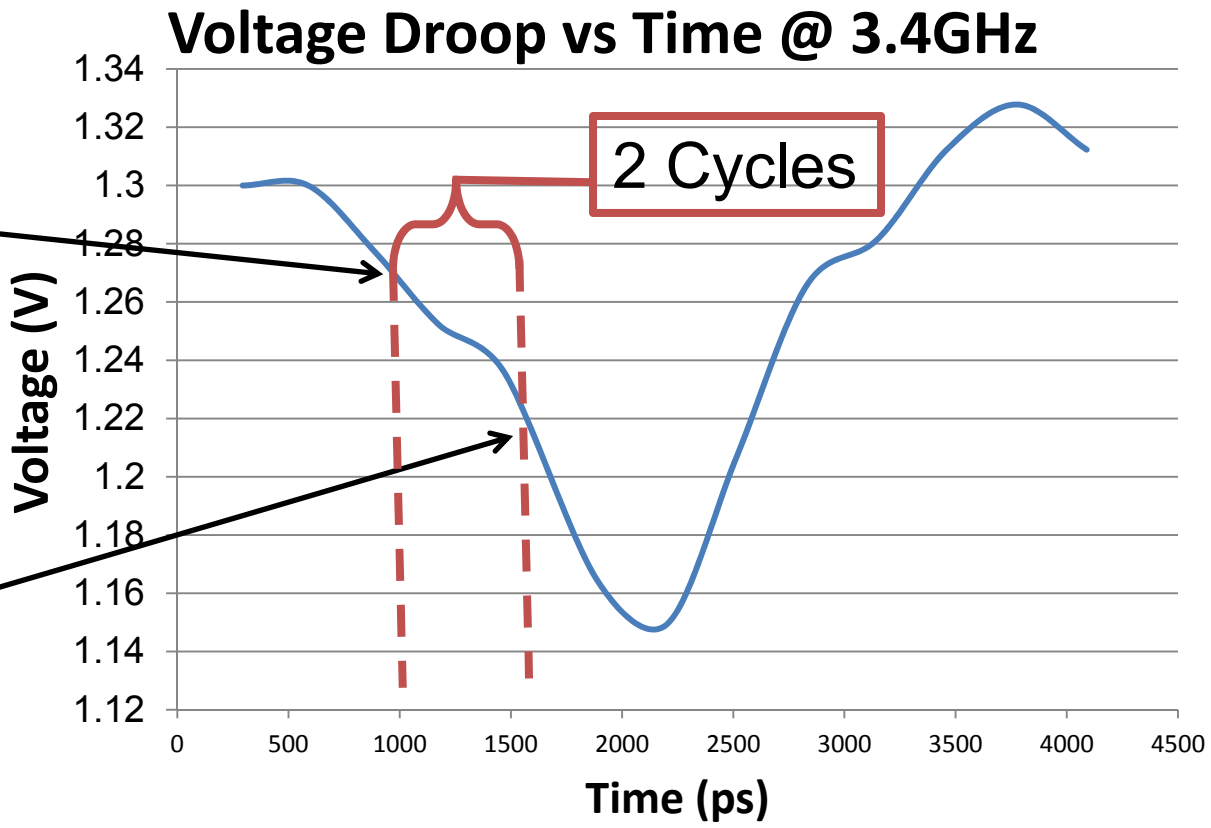
- To pick the clock:
  - The phase and phase enable are “and-ed”
  - Then all the enabled phases are “or-ed”
- Care has to be taken when routing the phases to prevent clock distortion

Si Results – Reaction Time

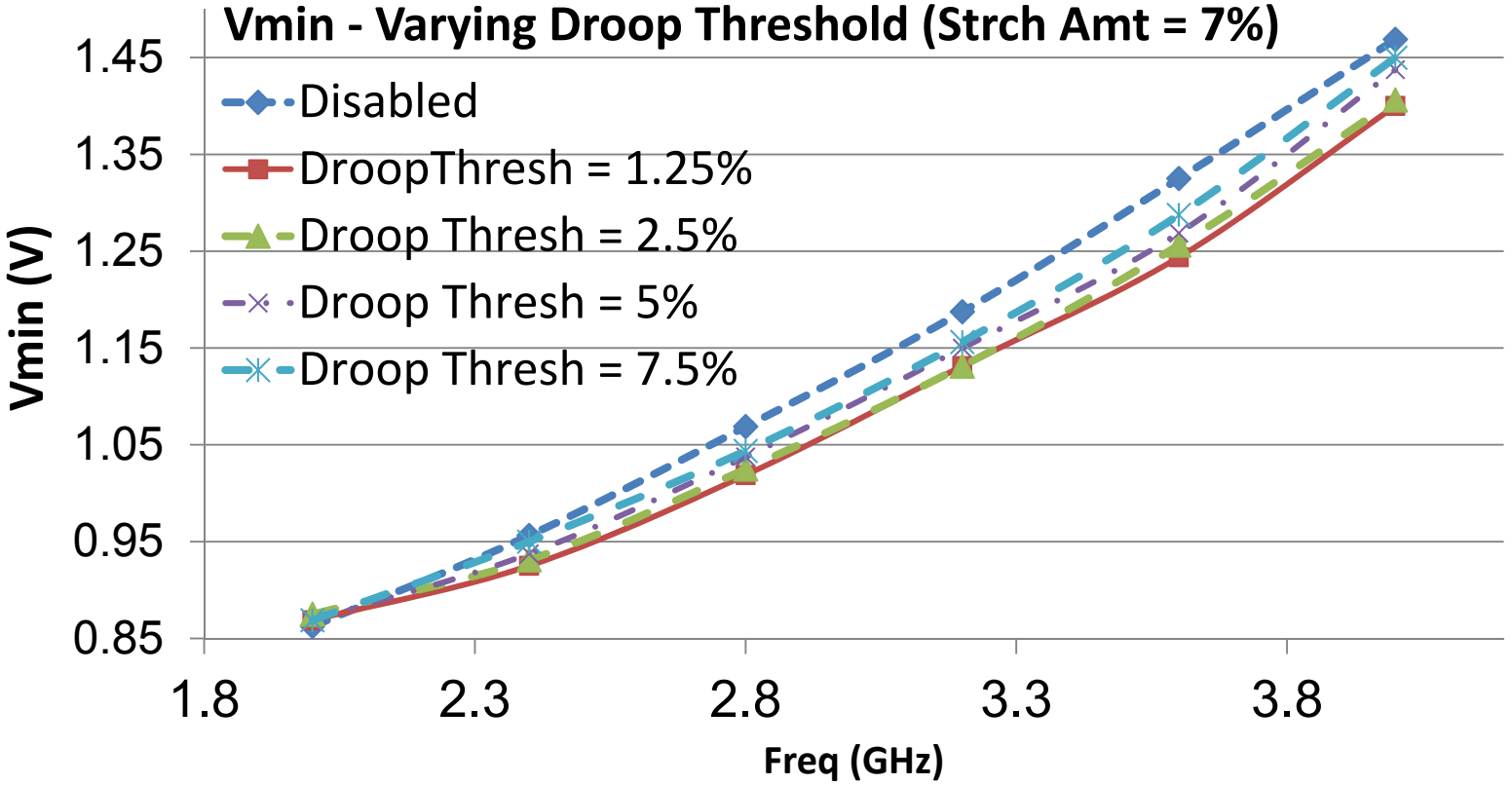
- Reaction time for droop detection can be as fast as 1 cycle
- The clock slows as fast as 2 cycles after droop is detected

DroopDetected

Clock frequency reduced after DroopDetected is synchronized into the clock generator domain

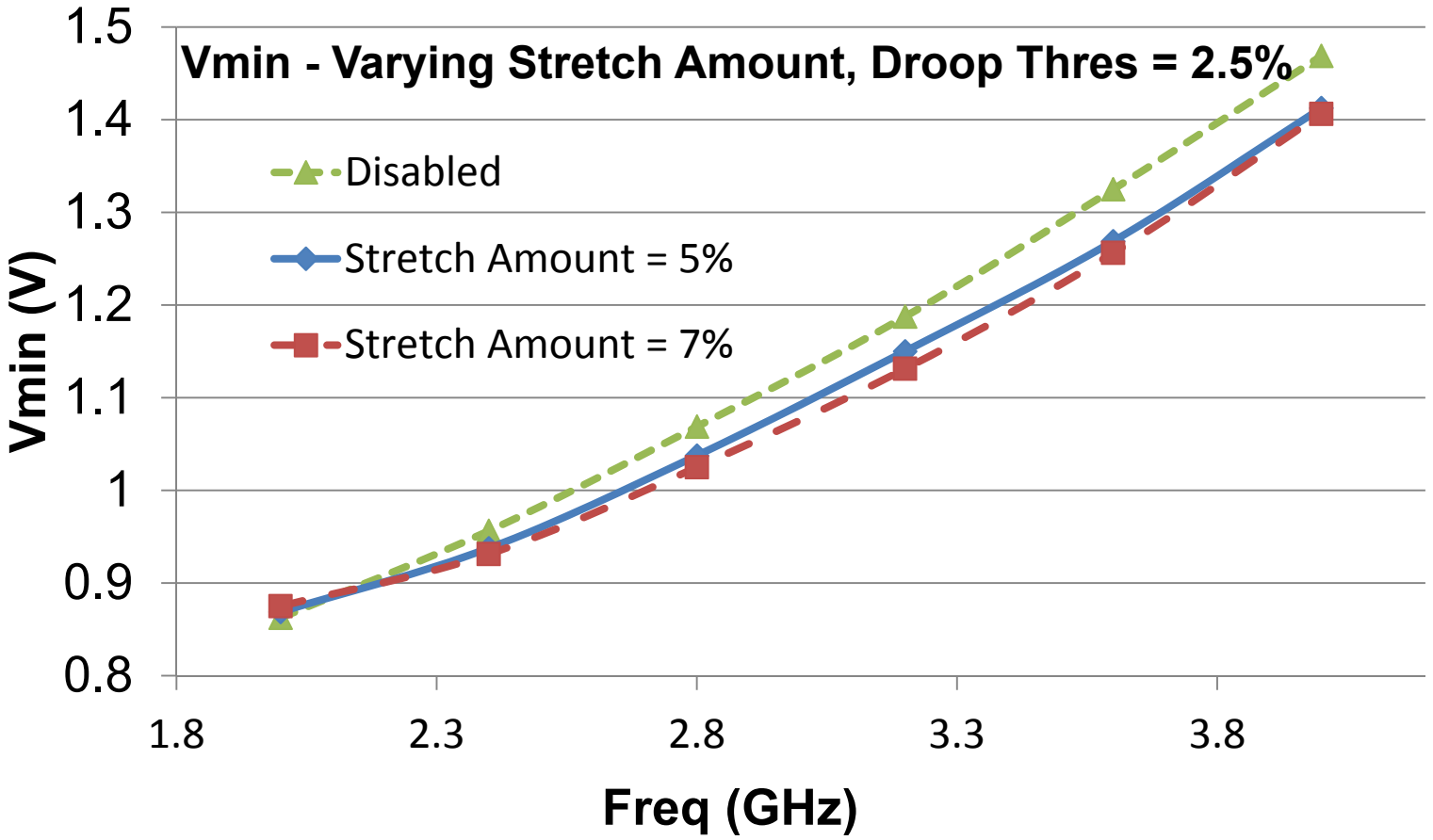


Si Results – Droop Threshold



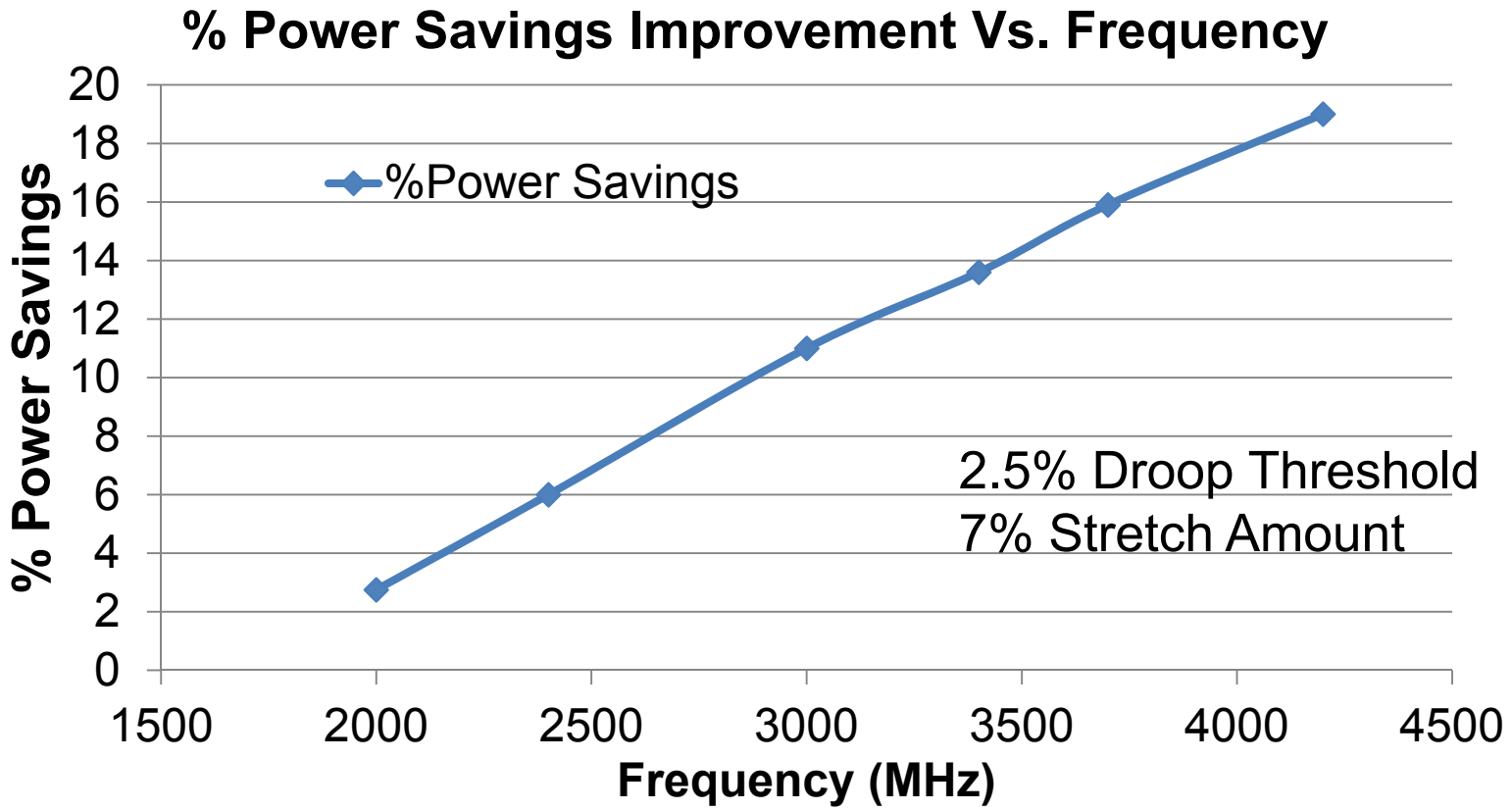
- Vmin for a given frequency decreases as the droop threshold is decreased
- Little benefit was observed between 2.5% and 1.25% of supply voltage

Si Results – Stretch Amount



- Vmin for a given frequency decreases as the stretch amount is increased

Si Results – Power Savings



- With production settings adaptive clocking decreases Vmin up to 90mV
- The resulting voltage reduction to achieve the same operating frequency results in up to 19% power reduction

Si Data - Performance Impact

Average Frequency with 2.5% Droop Threshold and 7% Stretch Amount @ 3400MHz		
	Compute Unit 0	Compute Unit 1
3DMark	3398 MHz	3396 MHz
CineBench	3397 MHz	3394 MHz
iTunes_acc	3396 MHz	3393 MHz
Itunes_mp3	3396 MHz	3394 MHz
POVRay	3394 MHz	3391 MHz
WinRAR	3400 MHz	3400 MHz

- Most applications have very few droop events that trigger the adaptive clocking system
- With a base frequency of 3400MHz, the Steamroller adaptive clock system typically loses <0.2% of cycles in effective frequency

Conclusion

- Transient droop events are rare, but cause a large impact to power through increased voltage to get a desired frequency.
- An adaptive clocking system that can detect the transient droop event and react quickly enough can mitigate the impact of these effects by temporarily decreasing the clock frequency.
- An adaptive clocking system has been demonstrated that can reduce power up to 19% with little to no measurable impact to performance.

A Graphics Execution Core in 22nm CMOS Featuring Adaptive Clocking, Selective Boosting and State-Retentive Sleep

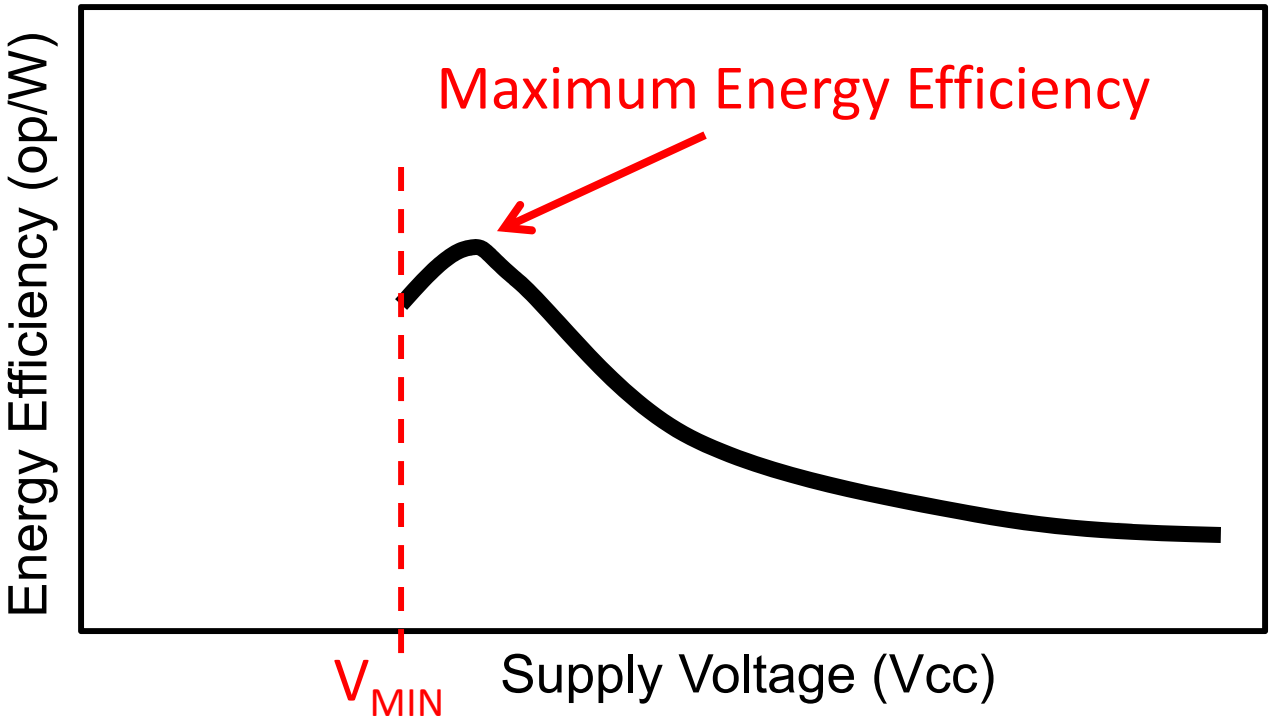
Carlos Tokunaga, Joseph F Ryan, Charles Augustine, Jaydeep P Kulkarni, Yi-Chun Shih, Stephen T Kim, Rinkle Jain, Keith Bowman, Arijit Raychowdhury, Muhammad M Khellah, James W Tschanz, Vivek De

Circuit Research Lab, Intel Corporation

Hillsboro, OR

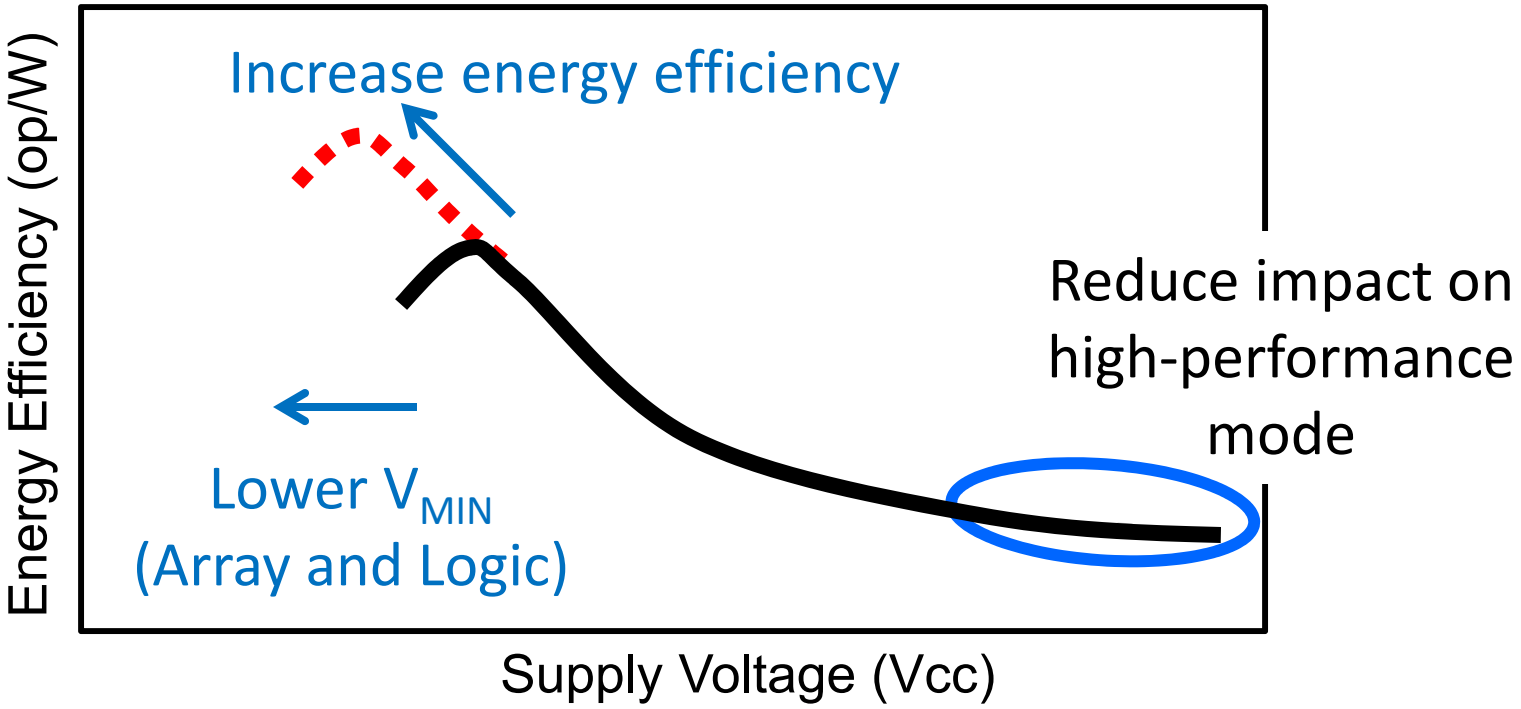
This research was, in part, funded by the U.S. Government (DARPA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Energy Efficiency in Graphics



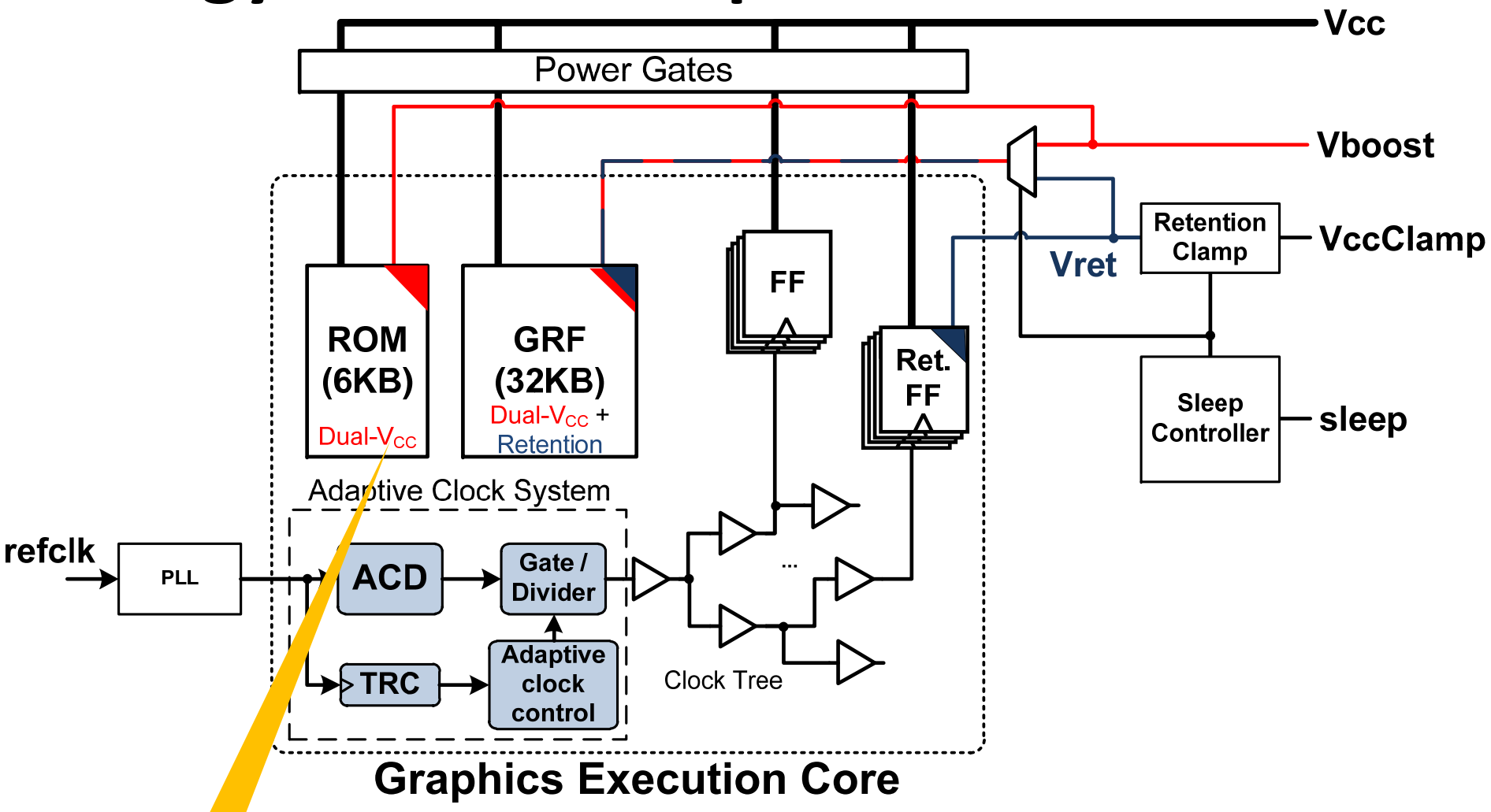
- Demand for high-performance graphics in power-constrained platforms (phones, tablets) is increasing.
- Goal: maximize energy efficiency without impacting high-performance mode

Energy Efficiency in Graphics



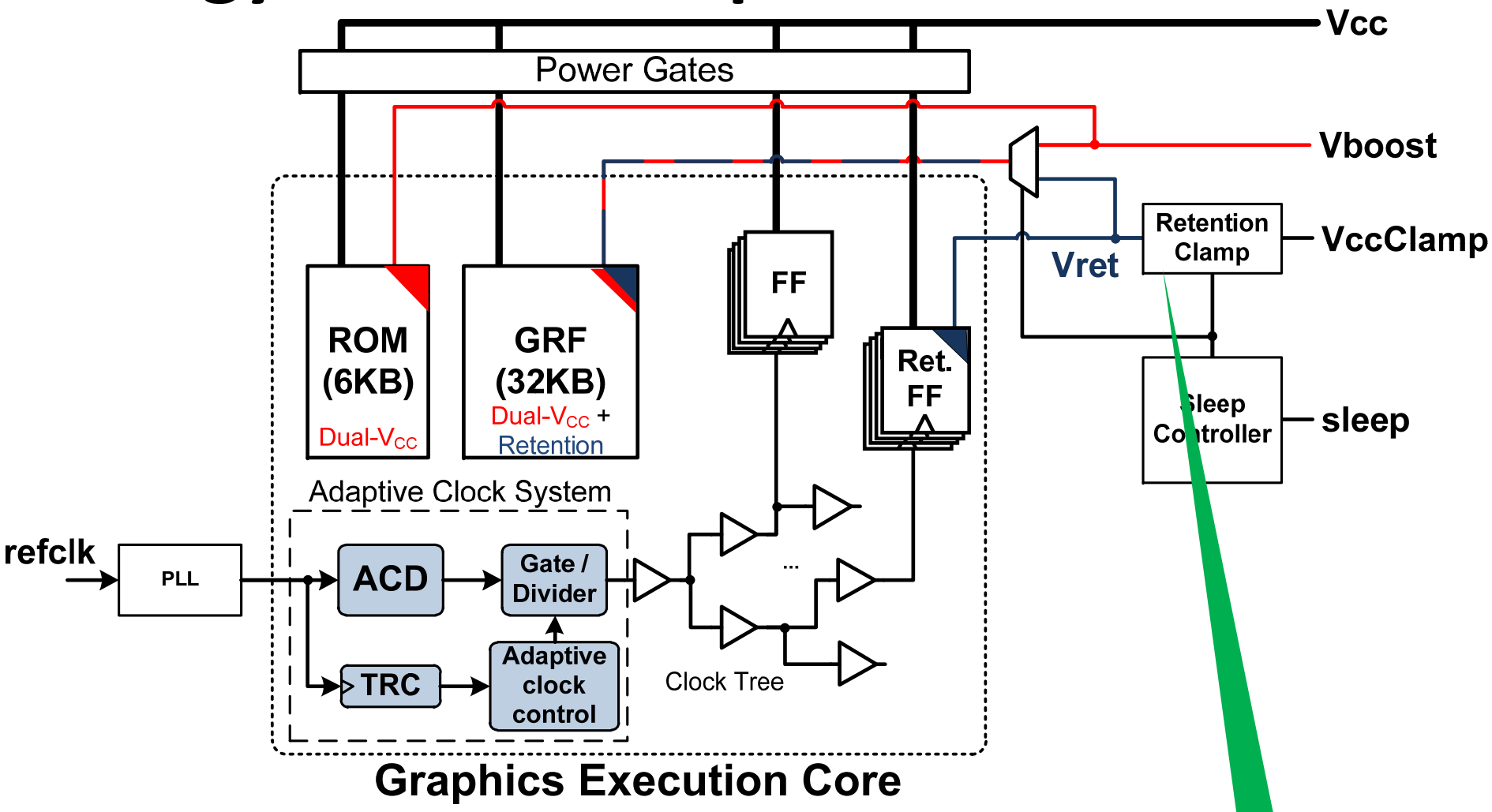
- Key techniques to improve V_{MIN} and increase energy efficiency
  - Array V_{MIN} reduction through selective boosting
  - Sleep leakage reduction through fast context save/restore
  - Voltage guardband reduction through adaptive clocking

Energy Efficient Graphics Execution Core

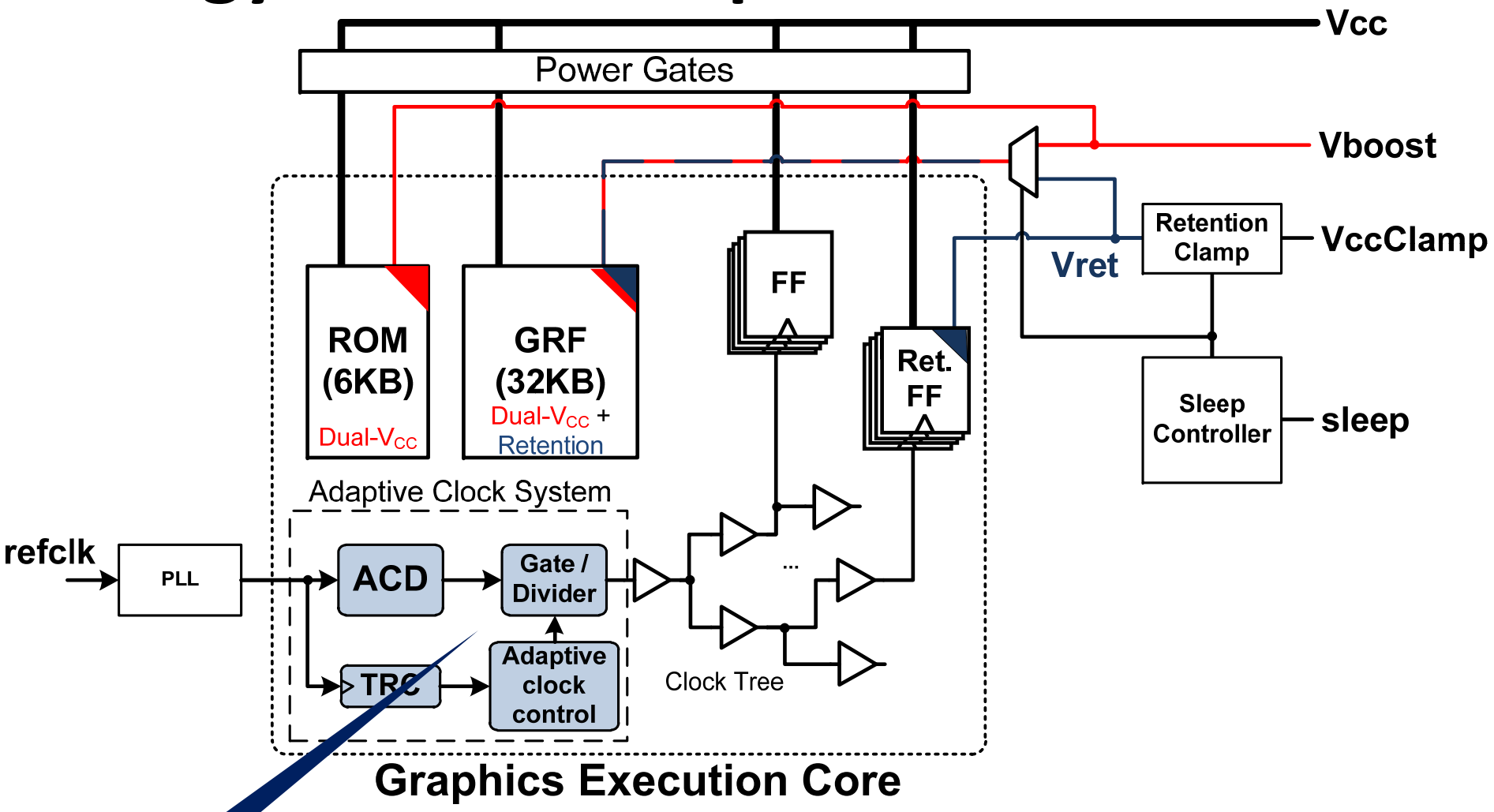


Dual-V_{CC} + Retention Register File
Dual-V_{CC} ROM

Energy Efficient Graphics Execution Core



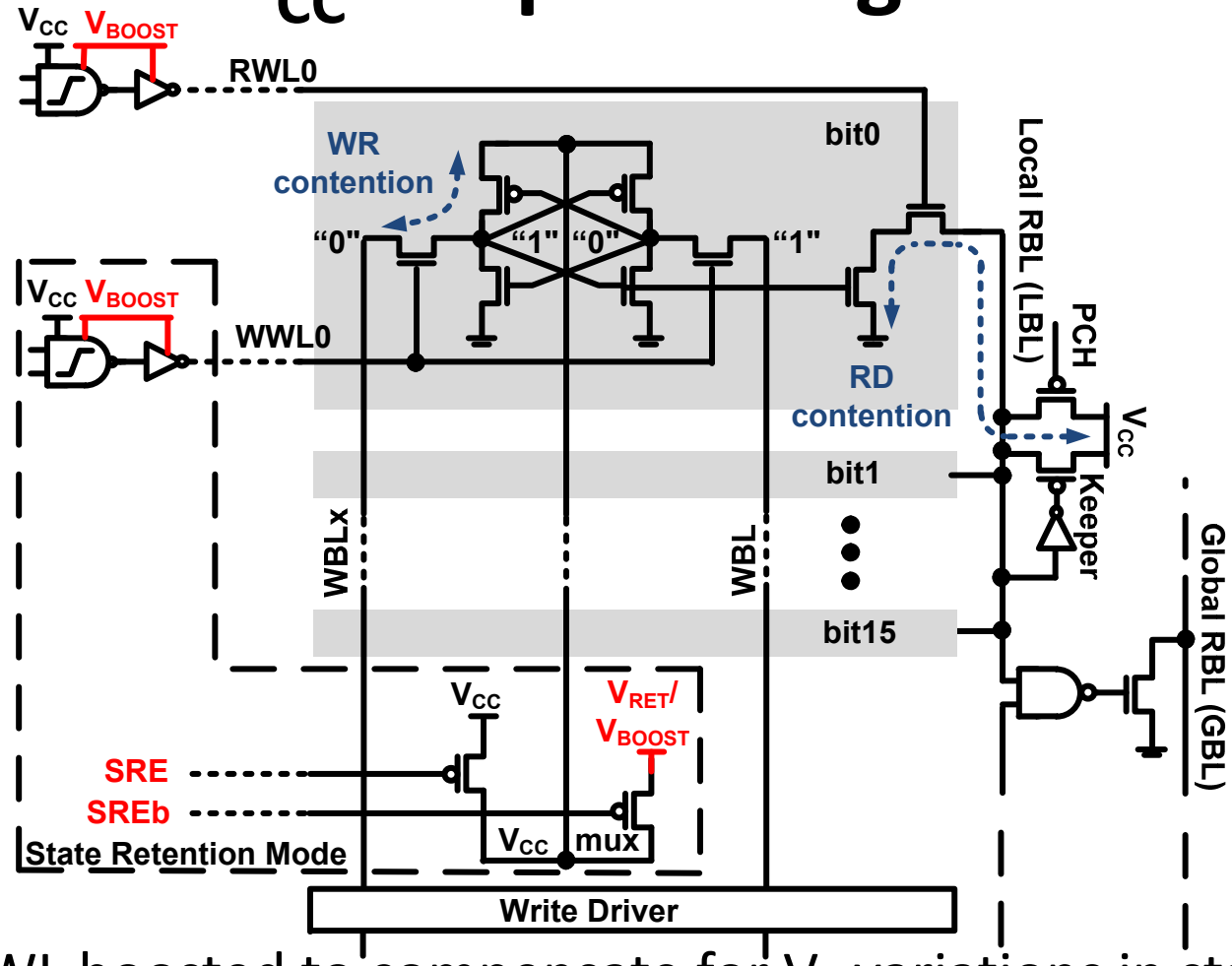
Energy Efficient Graphics Execution Core



Adaptive Clock System

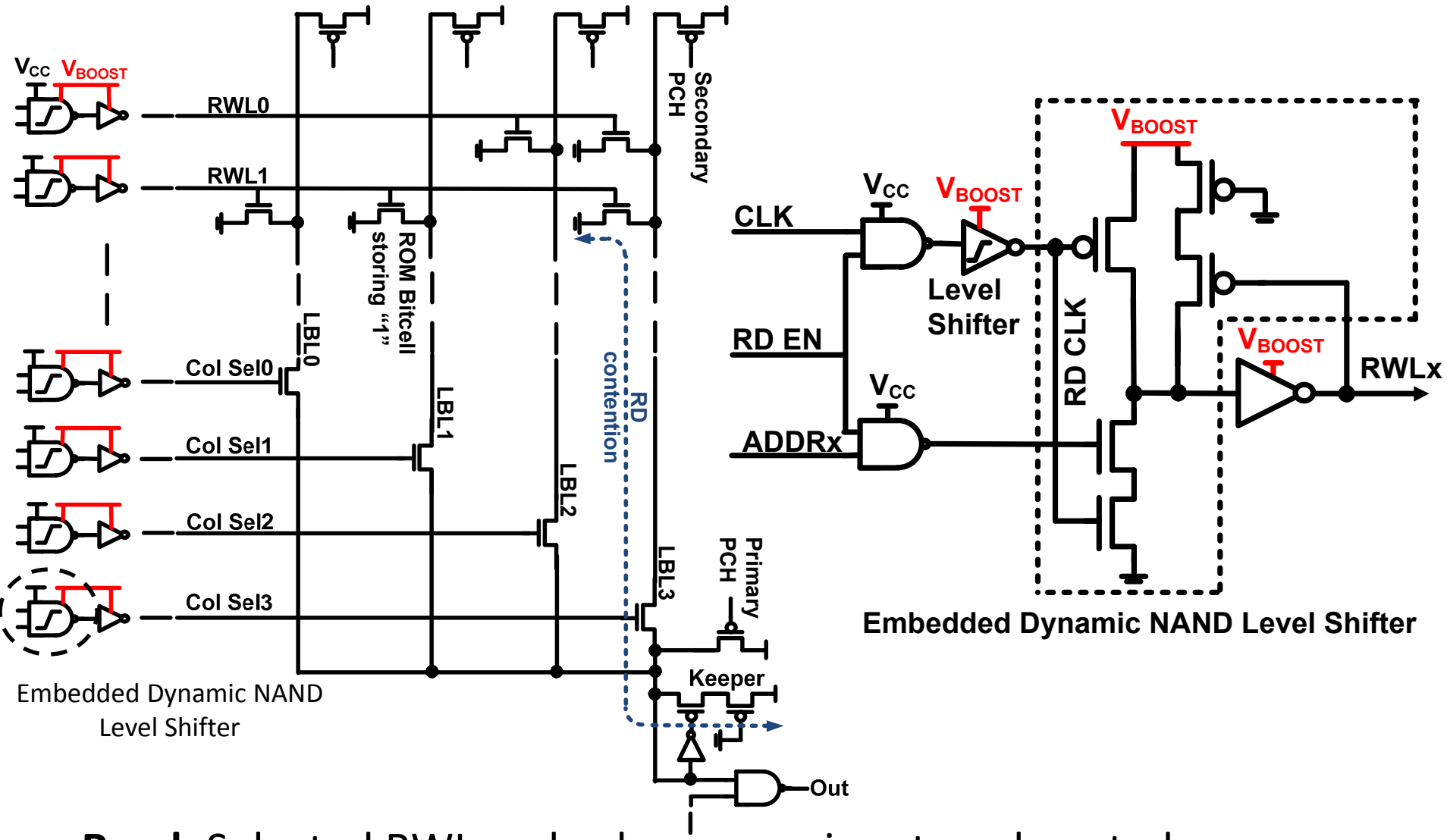
Tunable Replica Circuit + Adaptive Clock Distribution

Dual-V_{CC} Graphics Register File



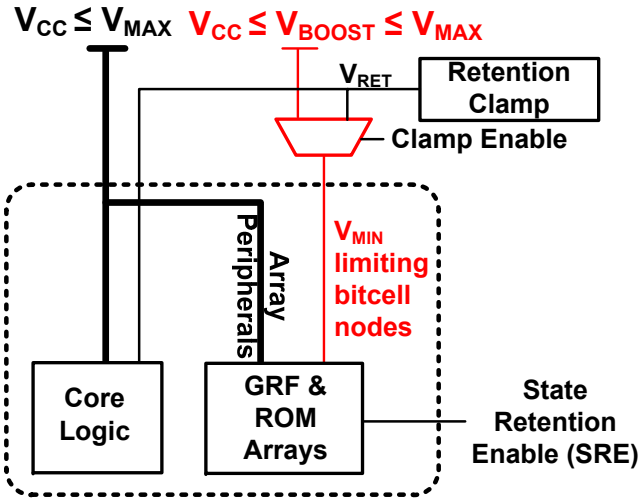
- **Read**: RWL boosted to compensate for V_T variations in stacked NMOS read port and/or PMOS keeper on local Read BL.
- **Write**: WWL boosted to mitigate contention between bitcell NMOS pass and PMOS pullup

Dual-V_{CC} ROM

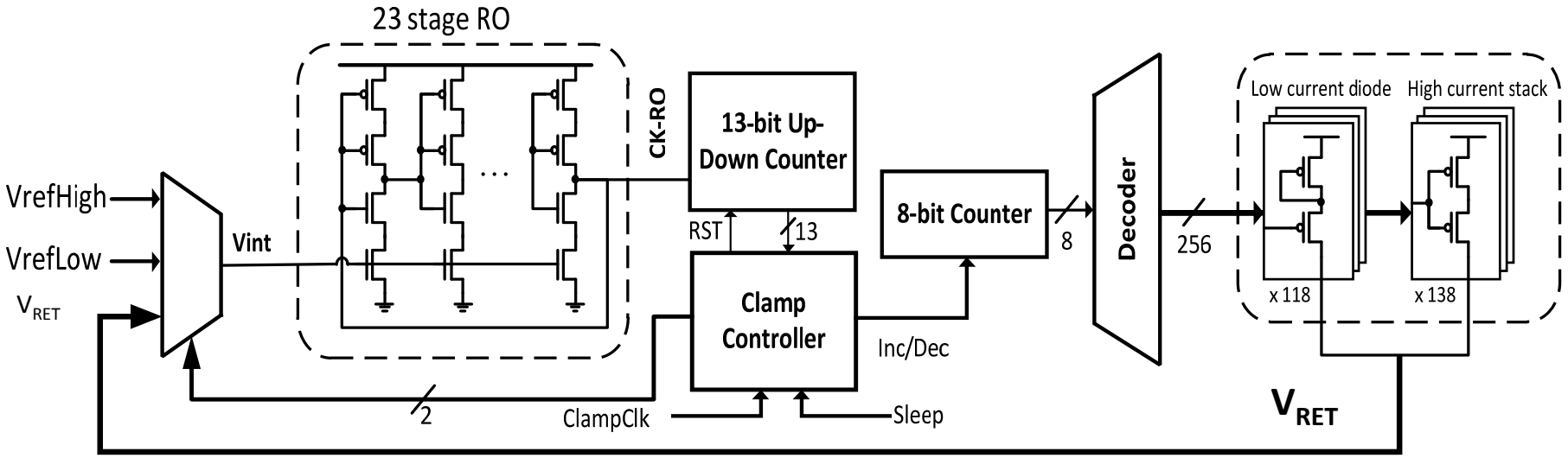


- **Read**: Selected RWL and column mux input are boosted using embedded dynamic level-shifting drivers

State Retention System

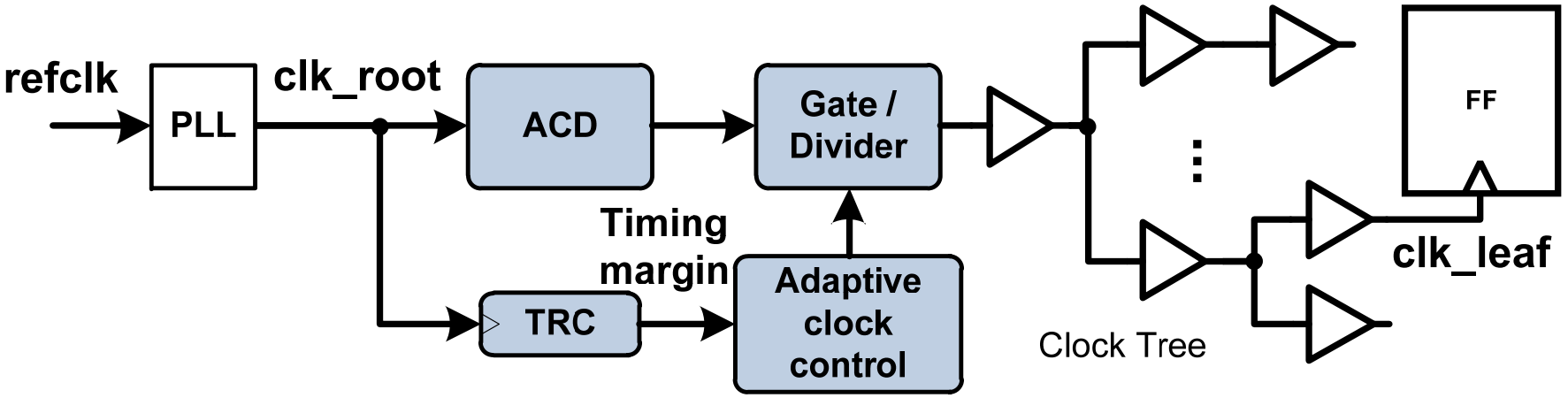


- State retention enables context save/restore for fast power gating
- GRF and sequentials connected to always-on supply V_{RET}
- All digital clamp reduces the always-on supply to the retention voltage limit

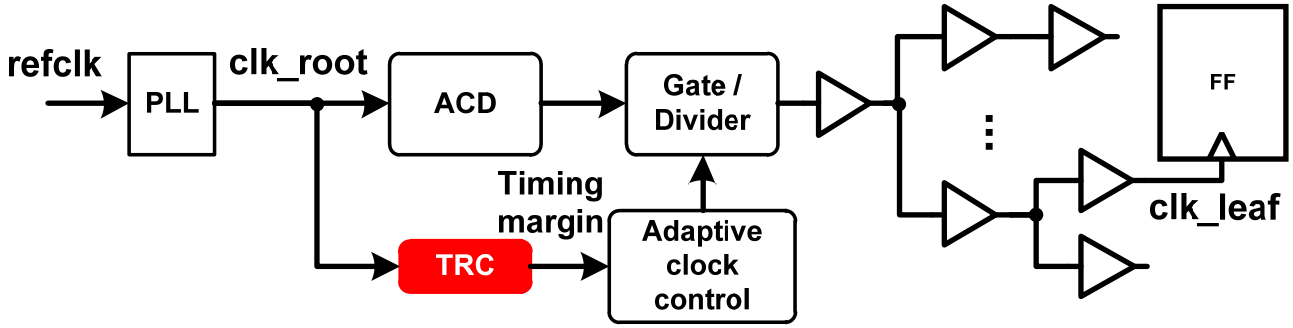


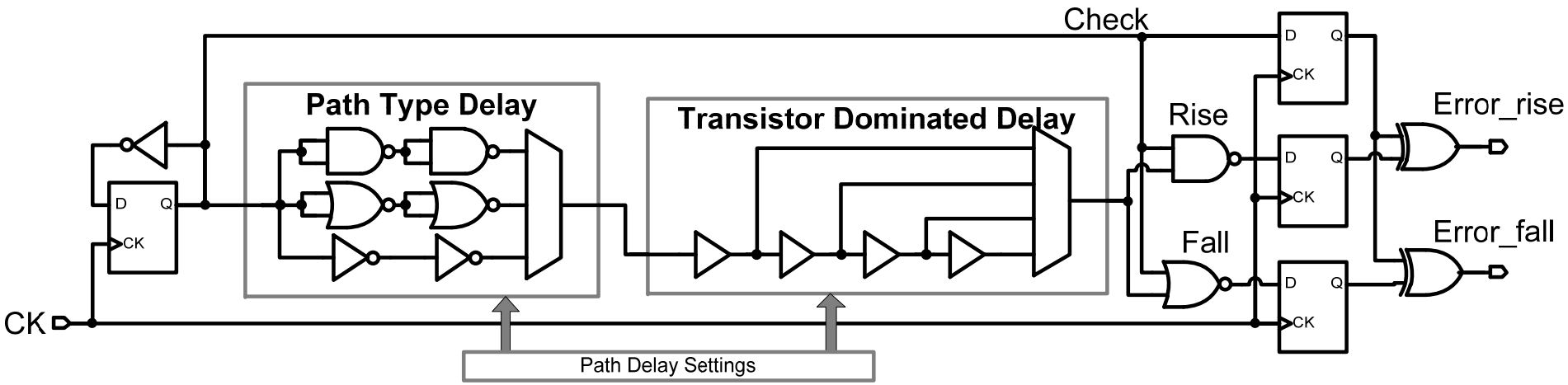
Adaptive Clock System

- Adaptive clocking technique mitigates the voltage guardband due to high-frequency (1st order) voltage droops
- Proactively gates or divides the core clock when a voltage droop is detected



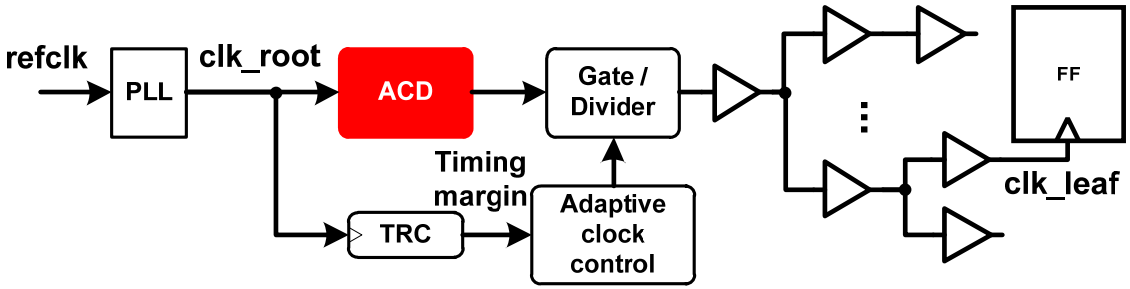
Tunable Replica Circuit

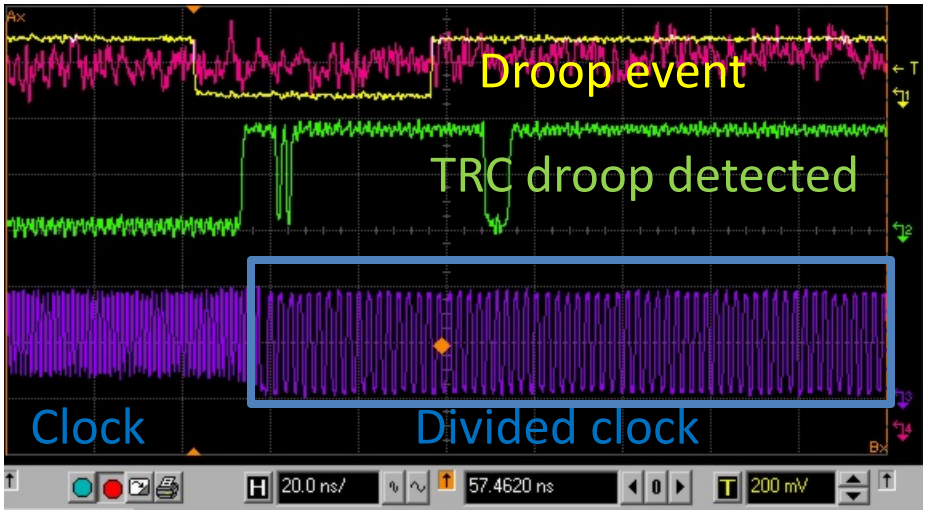




- Used as a voltage droop detector measuring critical-path timing margin every cycle
- Contains two programmable delay modules that are calibrated at test time to match critical-path delays

Adaptive Clock Distribution

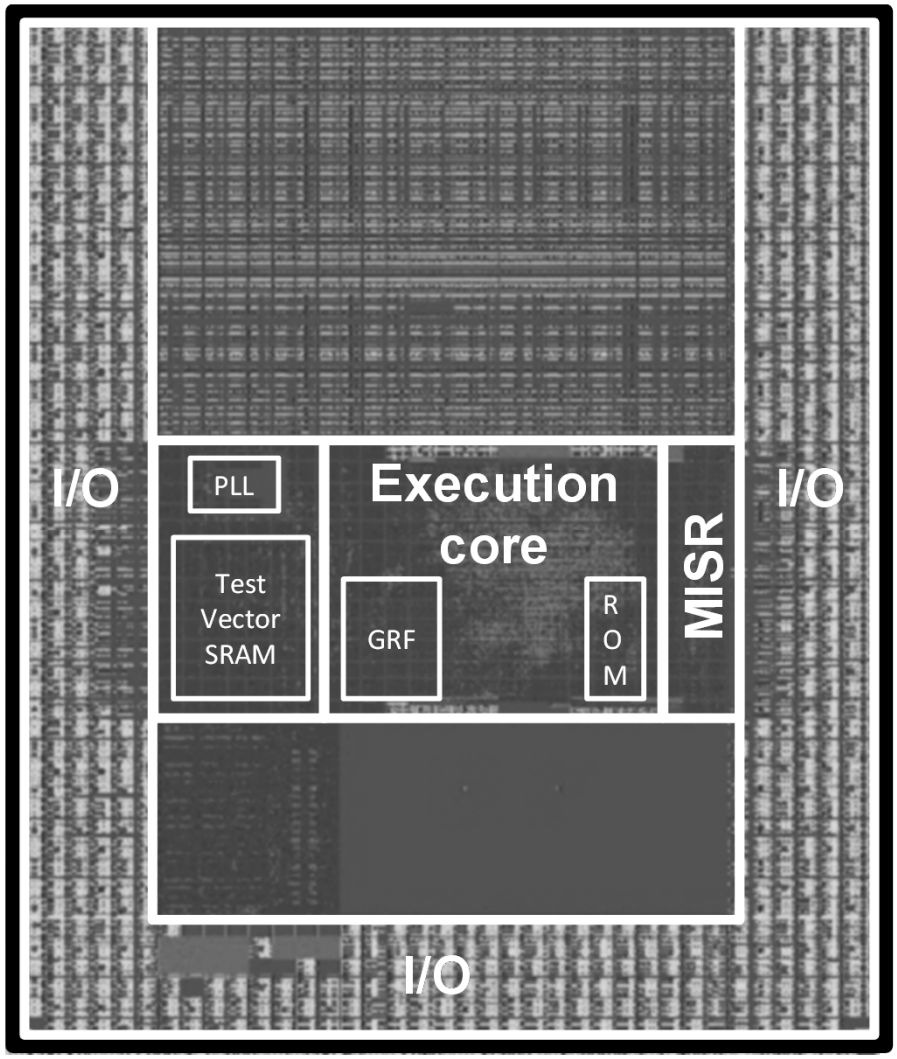


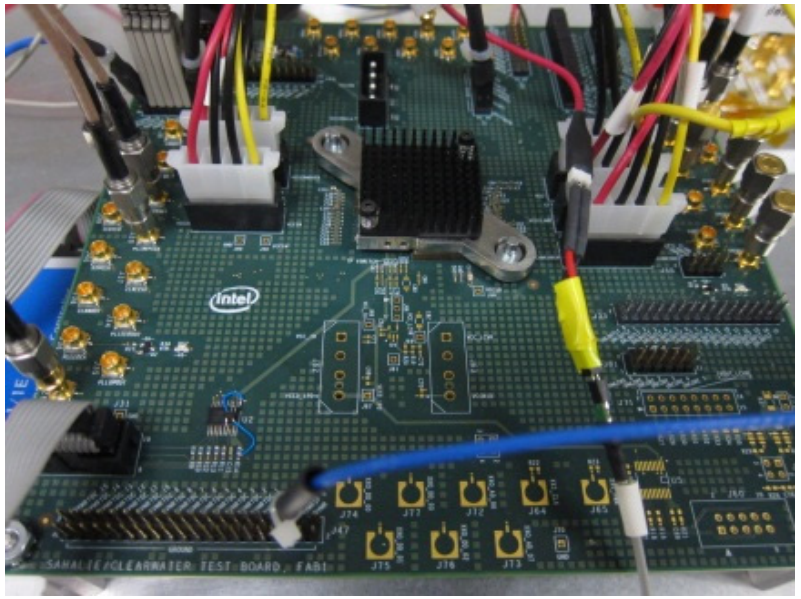


Guardband Reduction with ACD	
ACD Length (clock cycles)	F_{MAX} Recovery (%)
1	44
2	61
3	77
4	90
5	90

- Extends the delay and changes the delay sensitivity of the clock distribution to V_{CC}
- Prolongs clock stretching during a voltage droop providing time for the clock control to act

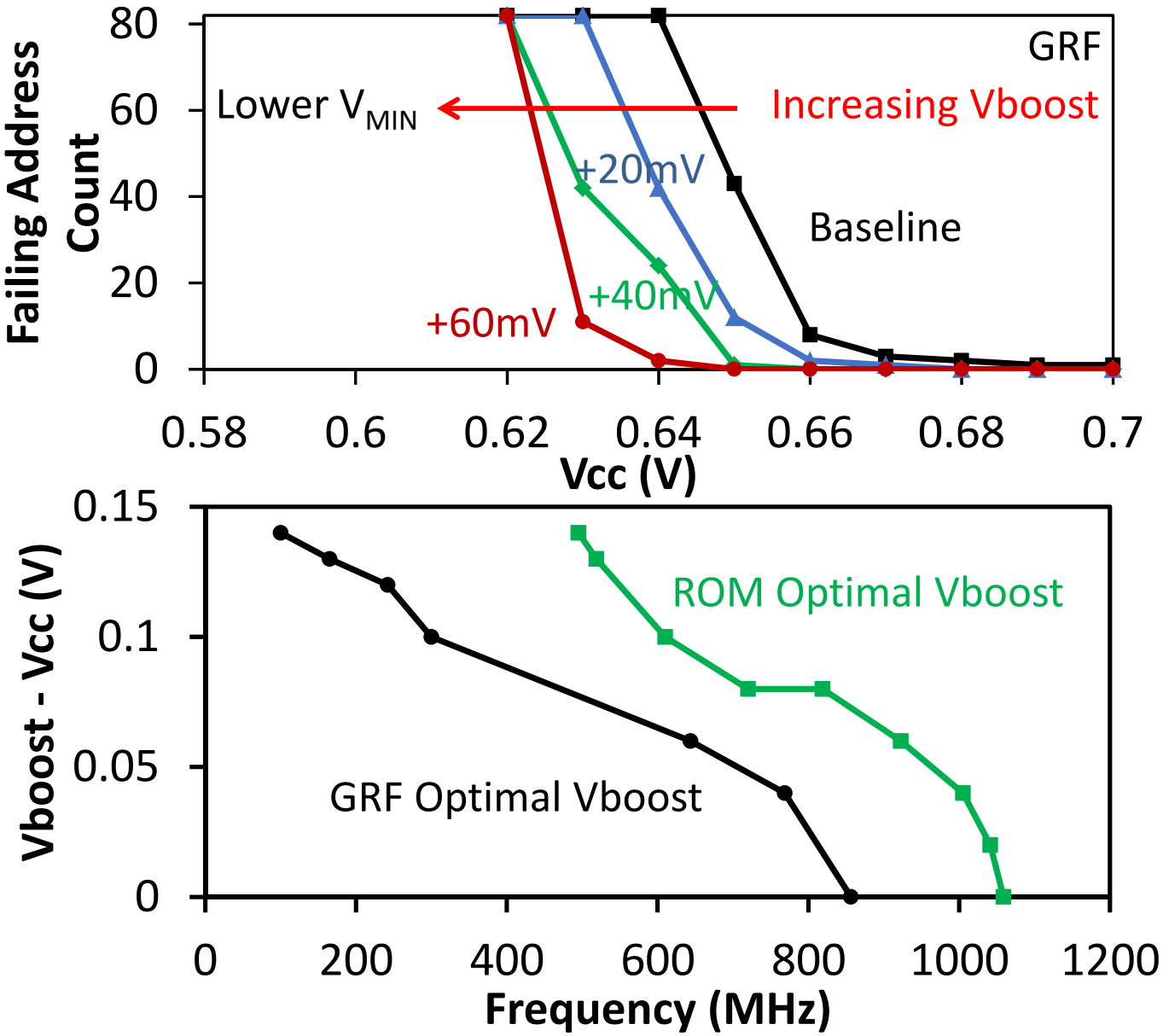
Testchip Micrograph and Design Details



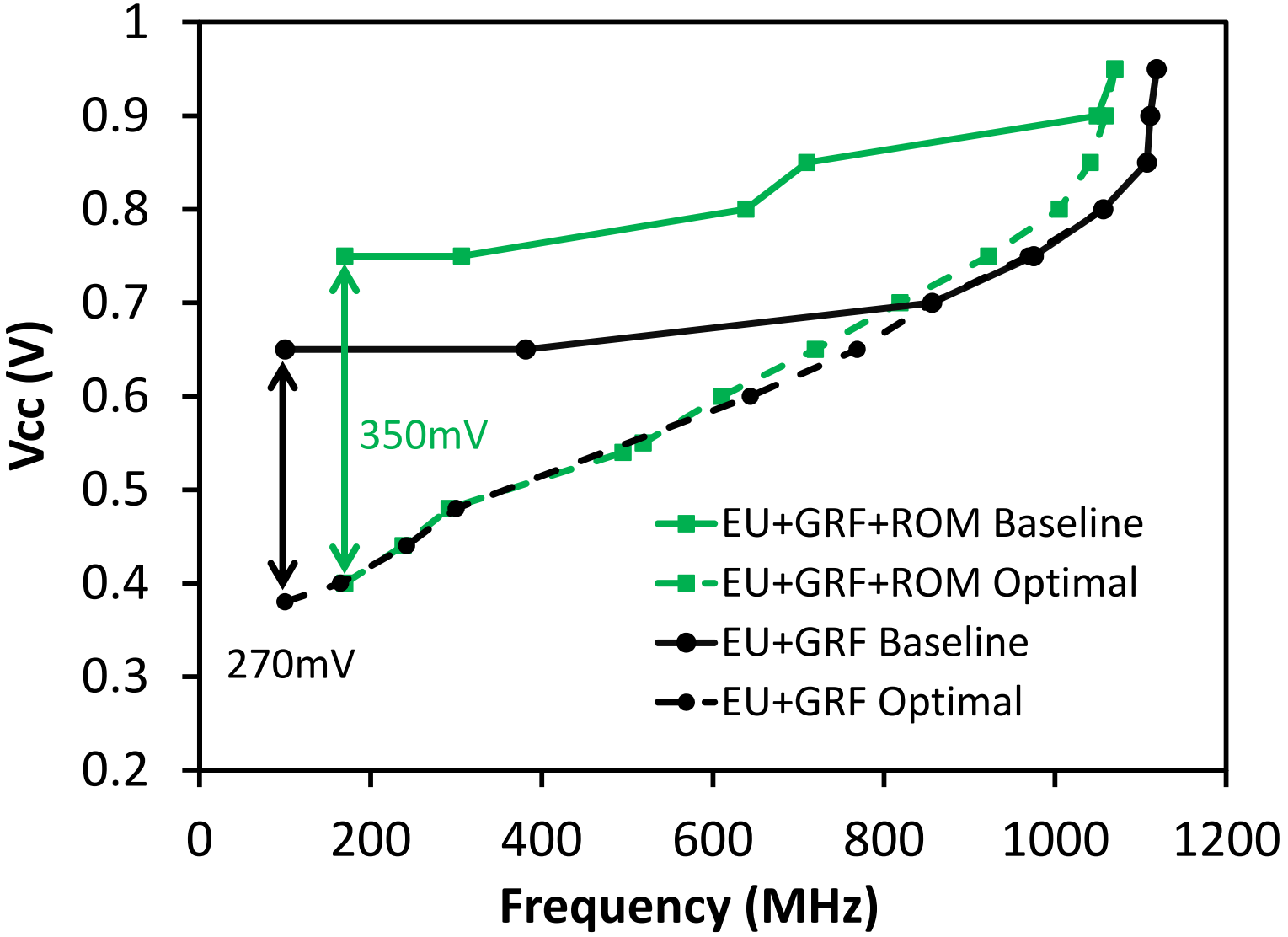


Technology	22nm, 9-metal layer tri-gate high-K/MG CMOS
Area: testchip die	4.0 x 5.8 mm^2
Area: core + test	2.6 x 1.3 mm^2
Core transistor count	22.8M
Target voltage, frequency	0.7V, 800MHz
Retention sequential count	14,411
Package	FCBGA13 951

Measured Impact of Selective Boosting

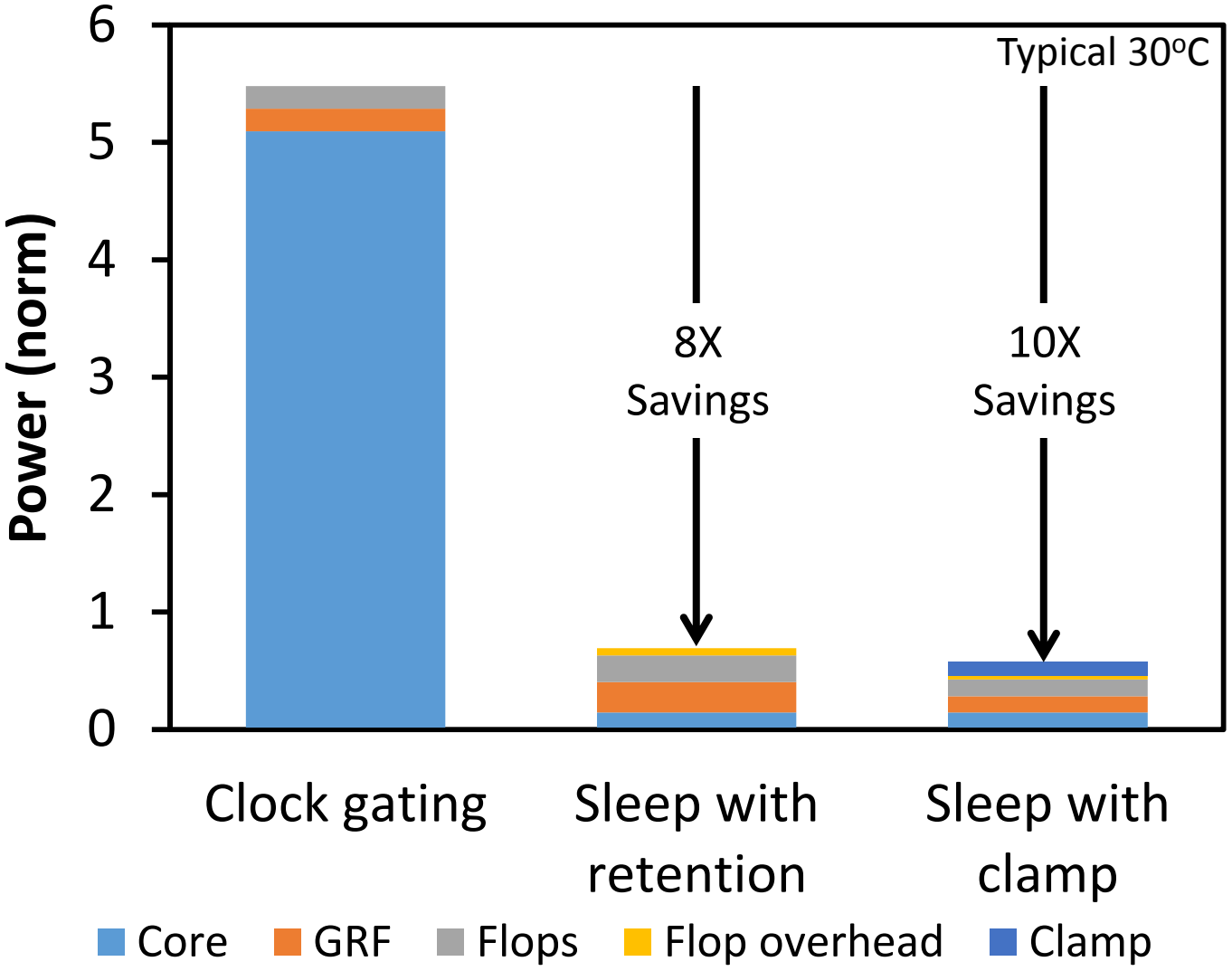


Measured V_{MIN} Reduction with Dual-V_{CC}



Measured V_{MIN} reduction of 270mV for RF and 350mV for ROM

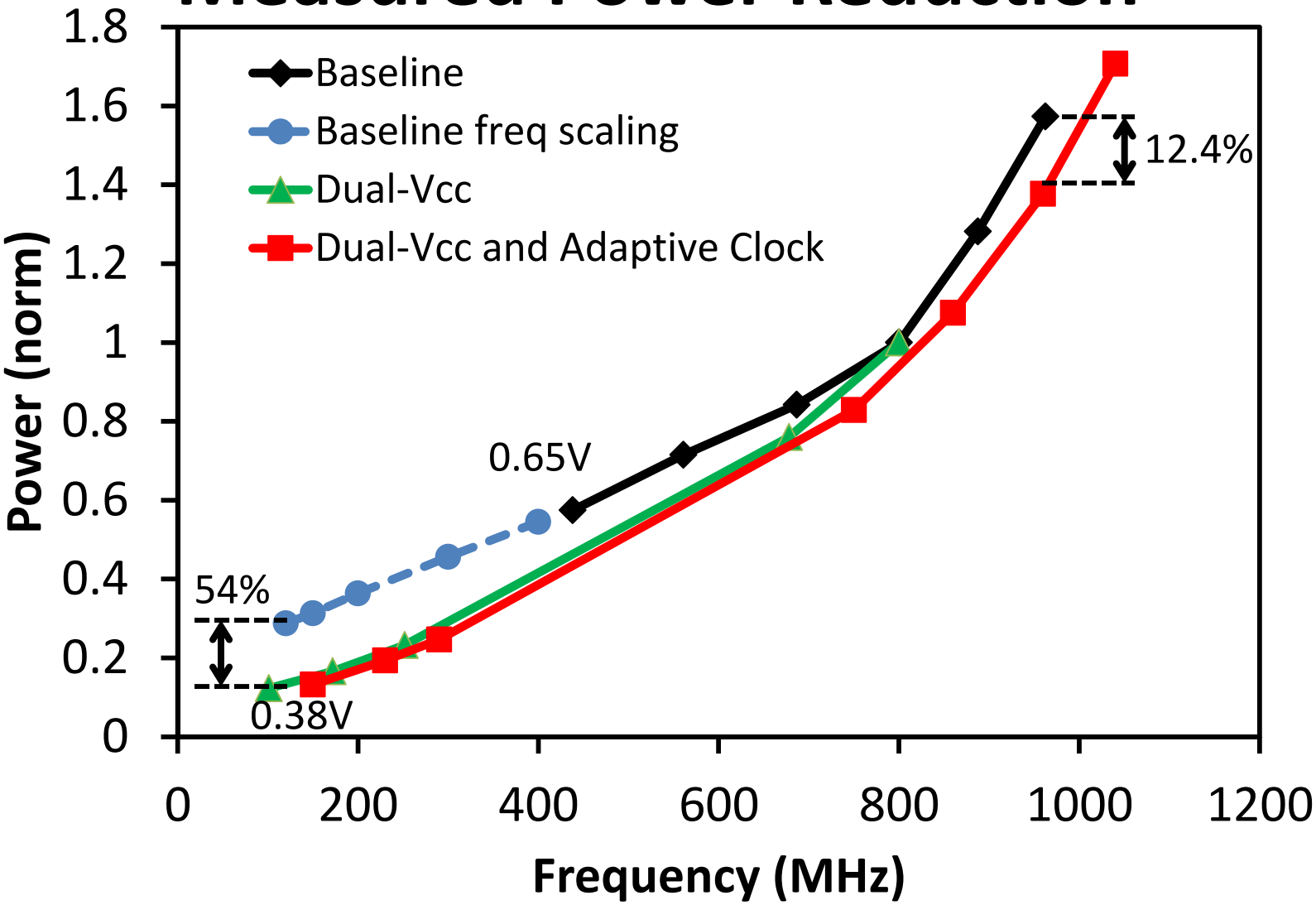
Measured Power Savings with Retention



Skew/Temp	Leakage Savings
Typical 30°C	10x
Typical 90°C	4x
Slow 30°C	20x
Slow 90°C	10x
Fast 30°C	10x
Fast 90°C	4x

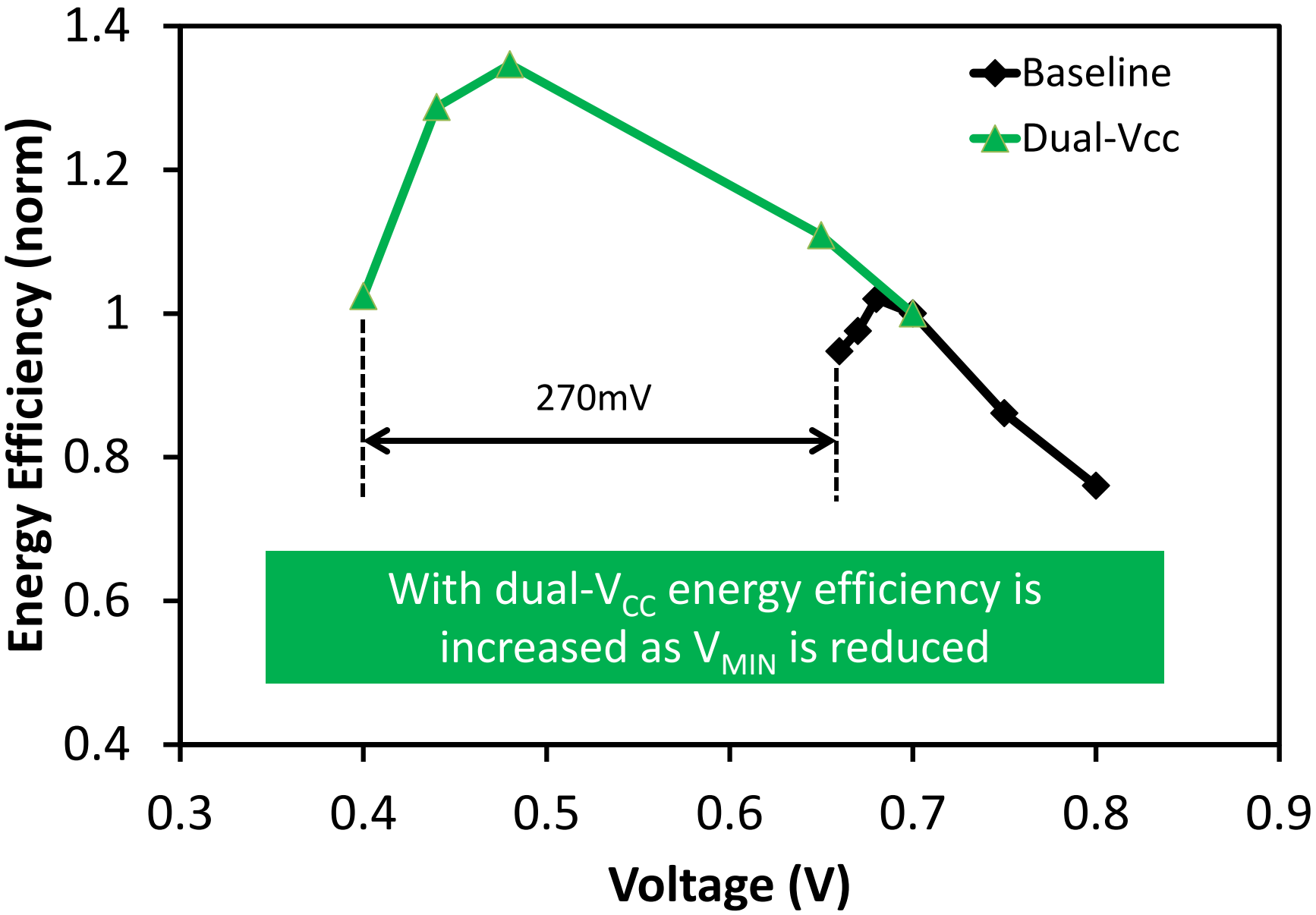
Retention clamp demonstrates leakage savings from 4X to 20X

Measured Power Reduction

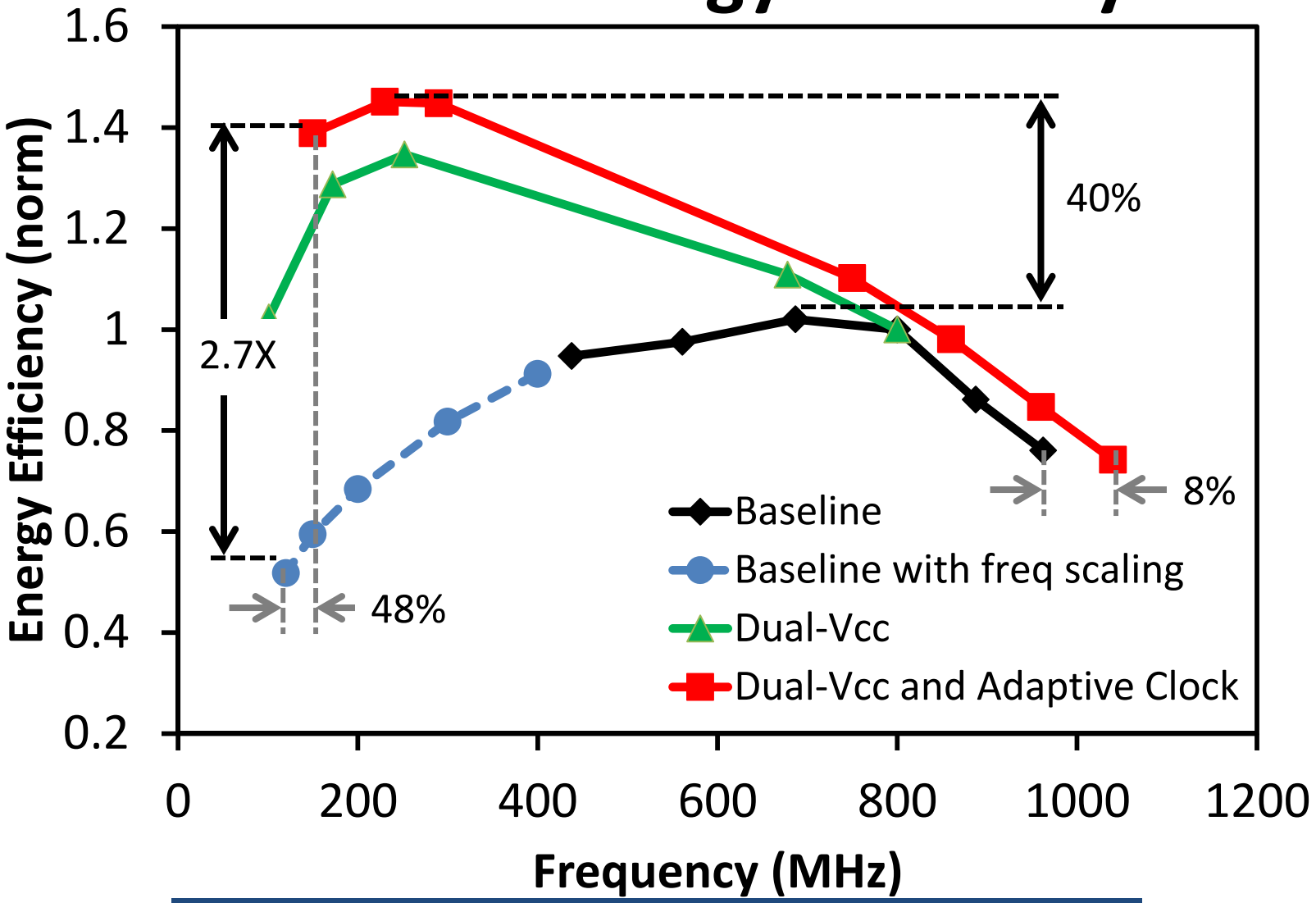


Power is reduced 54% at low frequency and 12.4% in high-performance mode

Measured Energy Efficiency



Measured Energy Efficiency



Measured energy efficiency increase up to 2.7X and 40% peak to peak

Summary

- Energy-efficient 22nm Graphics Execution Core integrates V_{MIN} reduction, state retention and adaptive clocking techniques.
- Dual-V_{CC} selective boosting provides V_{MIN} reduction of 270mV for the GRF, and up to 350mV for the ROM.
- Retention clamp demonstrates leakage savings in sleep mode from 4X to 20X.
- Adaptive clocking recovers up to 90% of the F_{MAX} guardband for high-frequency voltage droops
- Dual-V_{CC} and adaptive clocking together enable 1.4X improved energy efficiency and wide voltage operation range for the graphics execution core

A 3GHz 64-bit ARMv8 Processor in 40nm Bulk CMOS Technology

Alfred Yeung, Hamid Partovi, Qawi Harvard, Russ Homer, John Ngai, Luca Ravezzi, Matthew Ashcraft, Greg Favor

Applied Micro Circuit Corporation

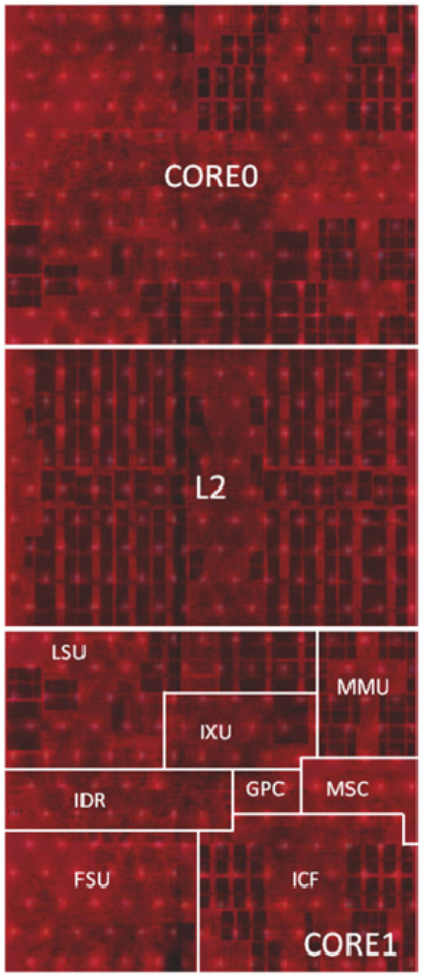
Agenda

- **Overview**
- Methodology
- Clock and power distribution
- Latching elements
- Custom macros
- Results and summary

Processor Overview

This presentation describes the design of a first generation 64-bit ARM v8 processor module (PMD) comprising:

- 2 identical CPU cores with a shared L2 cache
- 4-wide out-of-order superscalar micro-architecture
- Integer, FP and 128b SIMD engines
- Hardware virtualization support
- Hardware table-walk and nested page tables



Processor Overview

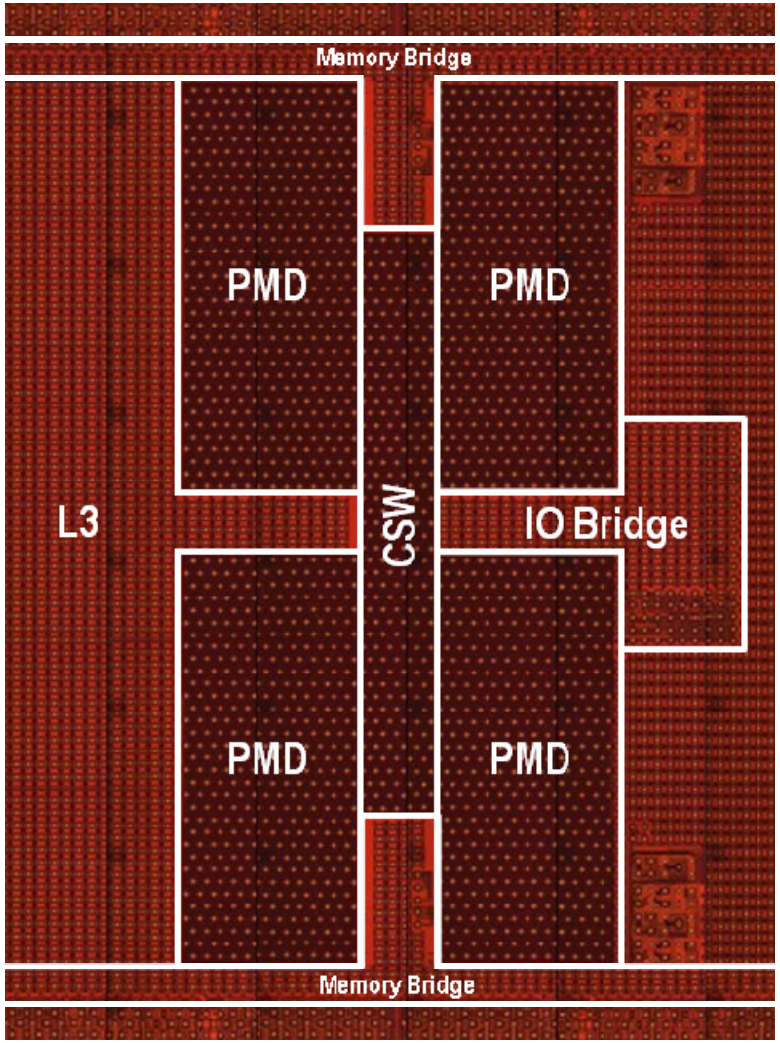
- Power management
  - Fine-grained clock gating
  - Software-hinted power states
  - Full DVFS
- Cache Hierarchy
  - Separate 32KB L1 instruction and data caches
  - Per-PMD 256KB L2 cache
  - Advanced hardware pre-fetch in L1 and L2

Processor complex and the SoC

- Implemented in a 40nm, 10 Metal, 3 V_T, bulk CMOS technology
- Configured with:
  - 4 PMDs
  - A shared 8MB L3 cache
  - 4 DRAM channels

arranged around a central switch

- Also integrates:
  - 10 Gb Ethernet
  - SATA 2/3
  - PCIe-3 and USB-3



Agenda

- Overview
- **Methodology**
- Clock and power distribution
- Latching elements
- Custom macros
- Results and summary

Design methodology

- Extensive design reuse
  - A small set of optimized custom memories
  - An enhanced standard cell library
- A robust flip-flop family capable of integrating logic
- Disallowed self-timing to reduce sensitivity to process variation
  - Memories use both edges of the clock, requiring firm control of duty-cycle
- Monotonic signaling to effect “soft” clock boundaries
- Clock-delayed dynamic domino circuits
- Circuit-based, tiered statistical analysis

Agenda

- Overview
- Methodology
- **Clock and power distribution**
- Latching elements
- Custom macros
- Results

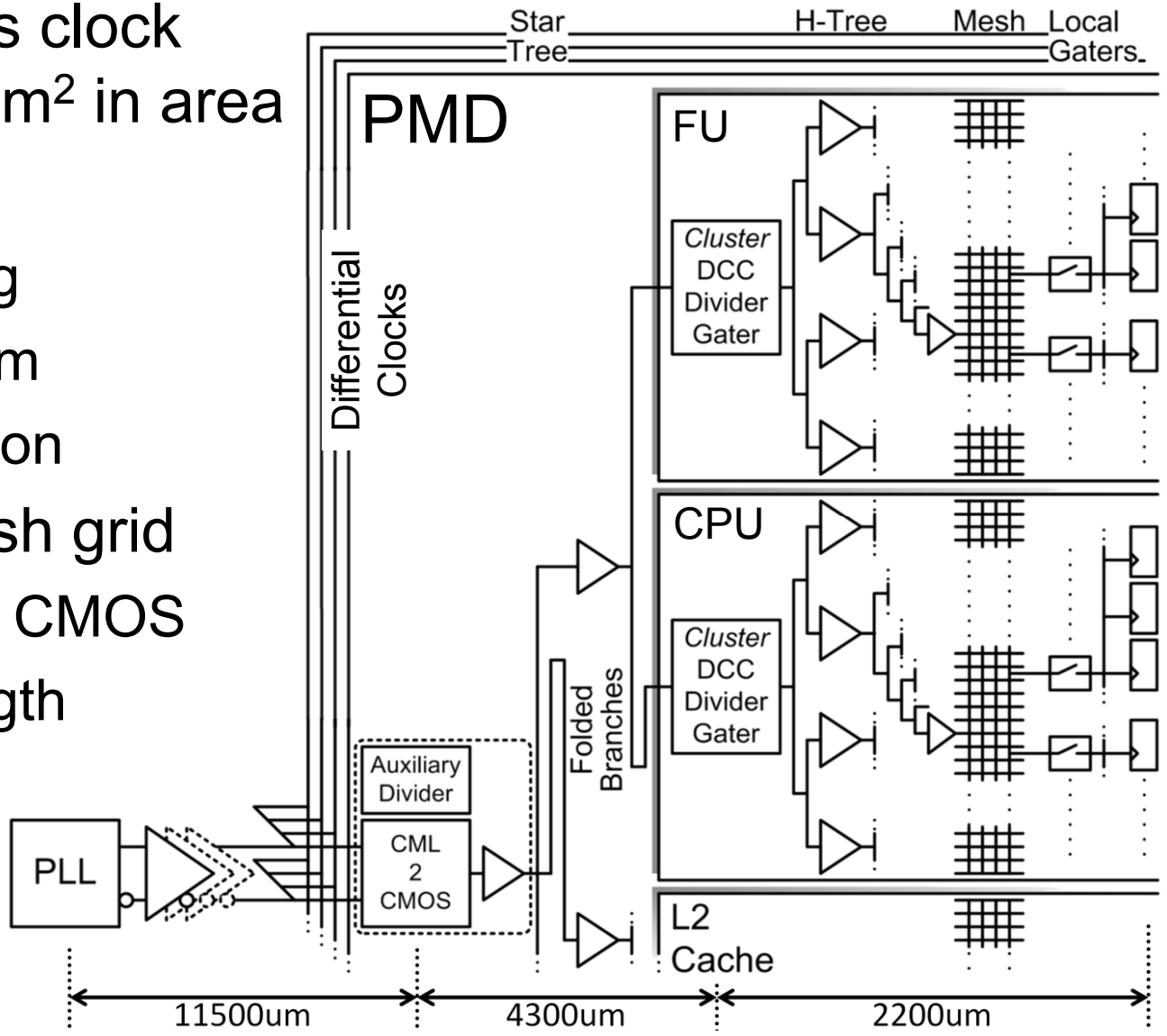
PLL

- Ring-oscillator based running at twice the processor operating frequency
- Its differential output drives front-end CML clock distribution to the 4 PMDs, L3 and the central switch
- Enables dynamic frequency-hopping for seamless DVFS operation
  - Frequency can be dynamically stepped down by 50% and back to full-rate in f_{REF} steps with negligible overshoot

Parameter	Value
Period Jitter (P|P)	< 1.0% t_{CK}
Absolute Jitter (RMS)	~ 1.5% t_{CK}
Frequency Hopping Overshoot	< 2.0% t_{CK}

Clock distribution

- Synchronous clock domain 15mm^2 in area
- Star tree
  - CML signaling
  - Spans 11.5mm
  - → C2C conversion
- H-tree + Mesh grid
  - Single-ended CMOS
  - 6.5mm in length



Duty-cycle correction and adjustment

- DCC circuit:
  - Detects, corrects and adjusts duty-cycle error accumulated in clock distribution network
  - Allows for distorting duty cycle by up to ±6% in 1% increments
  - Enables silicon debug of phase-based circuits by truncating critical phase
  - Extends critical phase; improves the operating frequency by 150MHz

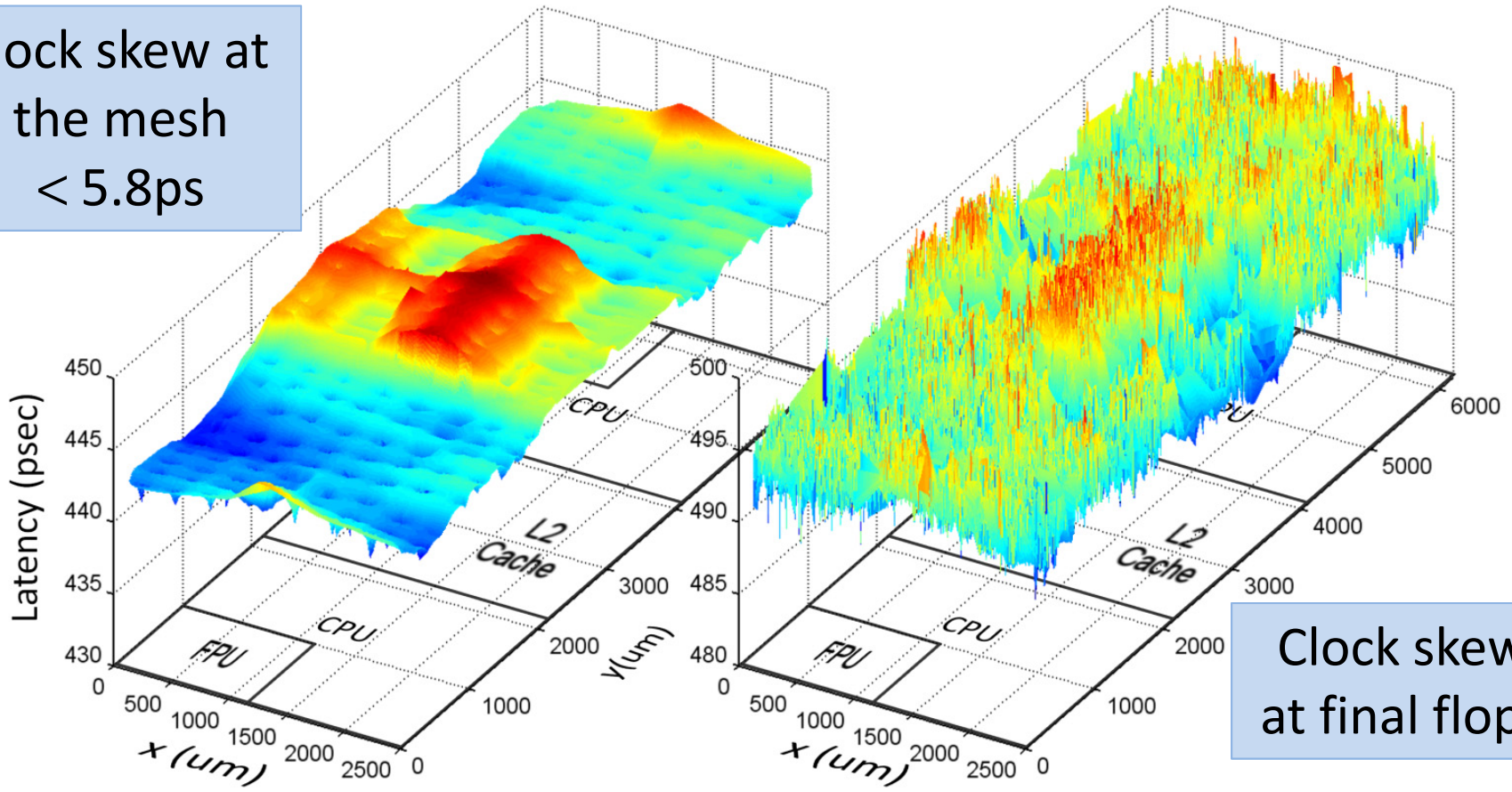
Number of passing cores

f_{CLK} (GHz)	BIST Test				Functional Test			
3.20	0	1	0	0	0	0	0	0
3.15	1	1	1	0	0	0	0	0
3.10	1	2	1	0	0	0	0	0
3.05	3	8	3	0	0	8	2	0
3.00	5	8	8	3	4	8	6	0
2.95	6	8	8	3	4	8	8	2
2.90	8	8	8	8	8	8	8	4
2.85	8	8	8	8	8	8	8	8
	0	+1	+2	+3	0	+1	+2	+3

Duty Cycle Offset (%)

PMD-wide clock skew and jitter

Clock skew at the mesh
< 5.8ps

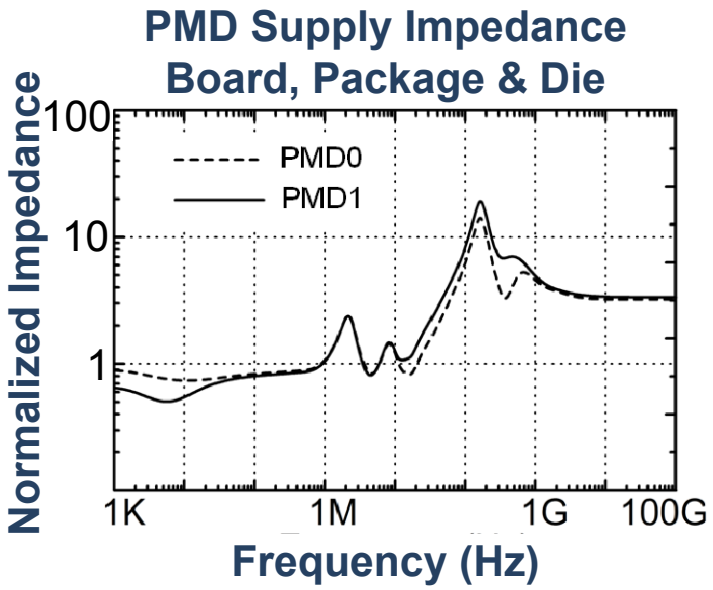


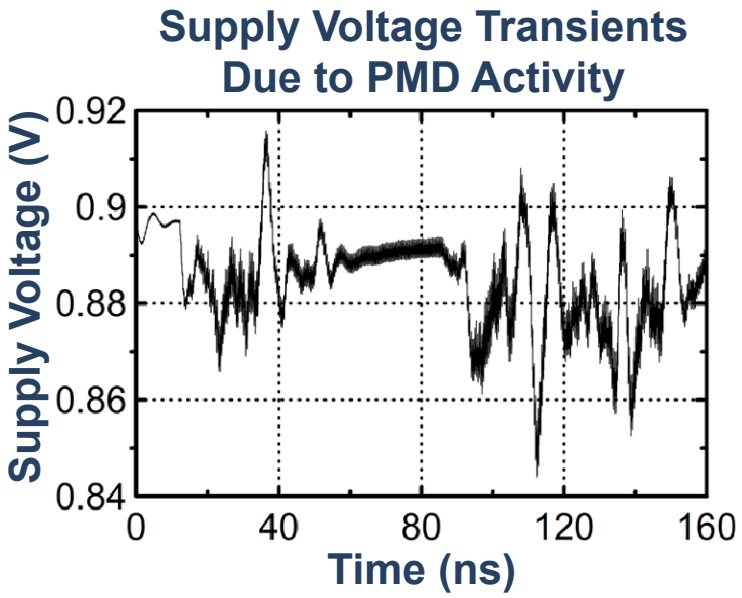
Clock skew at final flops

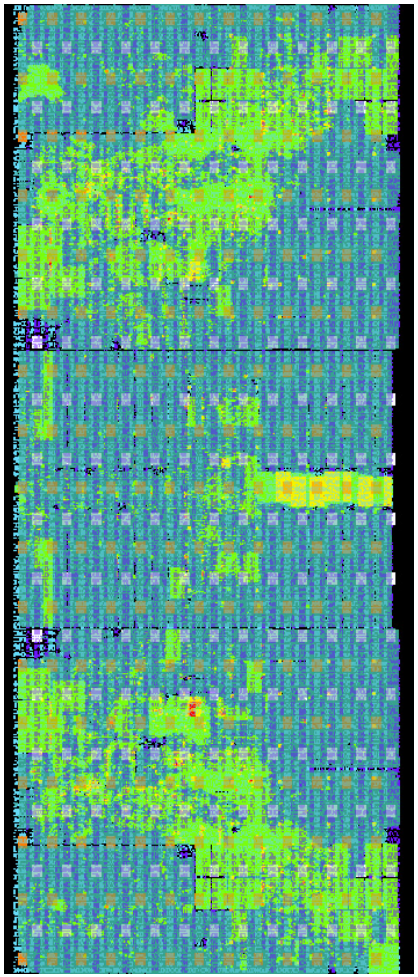
Skew and Jitter at the Flip-flops	Value
Clock Skew	< 13.7ps
Sensitivity to Supply Noise	< 0.8ps/mV$\|_{RMS}$

Power delivery

- Thick metal layers (M9/M10) provide 95% of low-resistance planar distribution
- Stacked Vias minimize vertical resistance
- Mesh based design ensures low impedance to Dcaps and minimizes variability





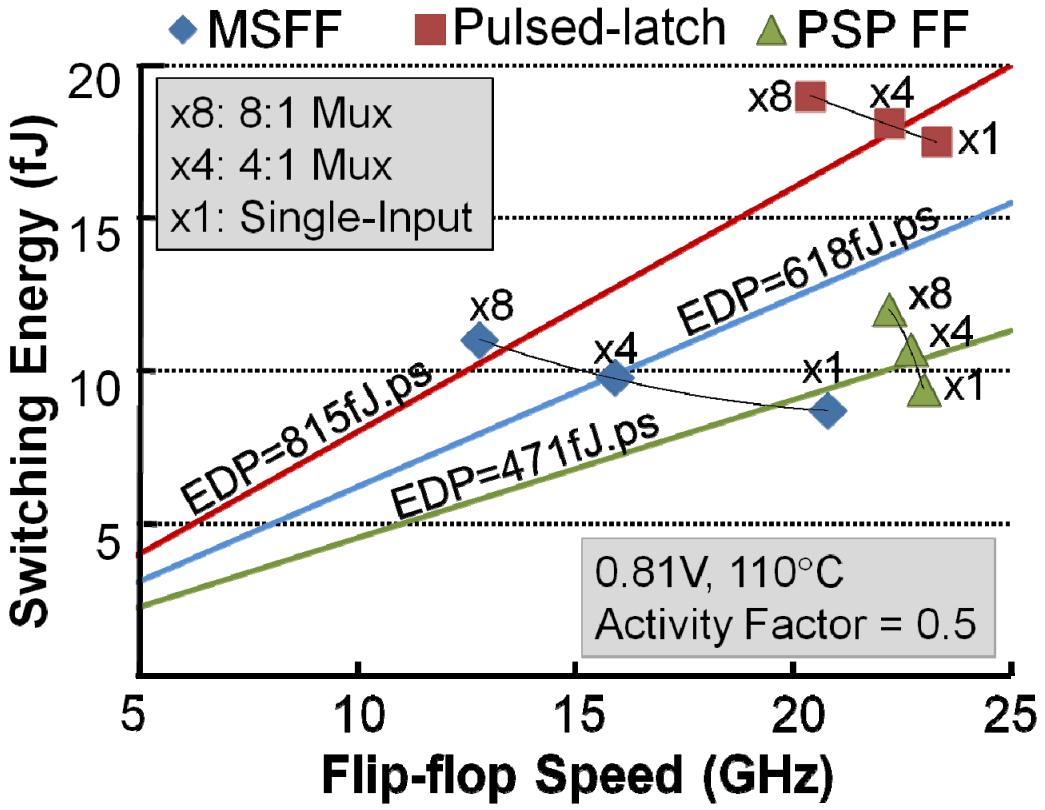


Agenda

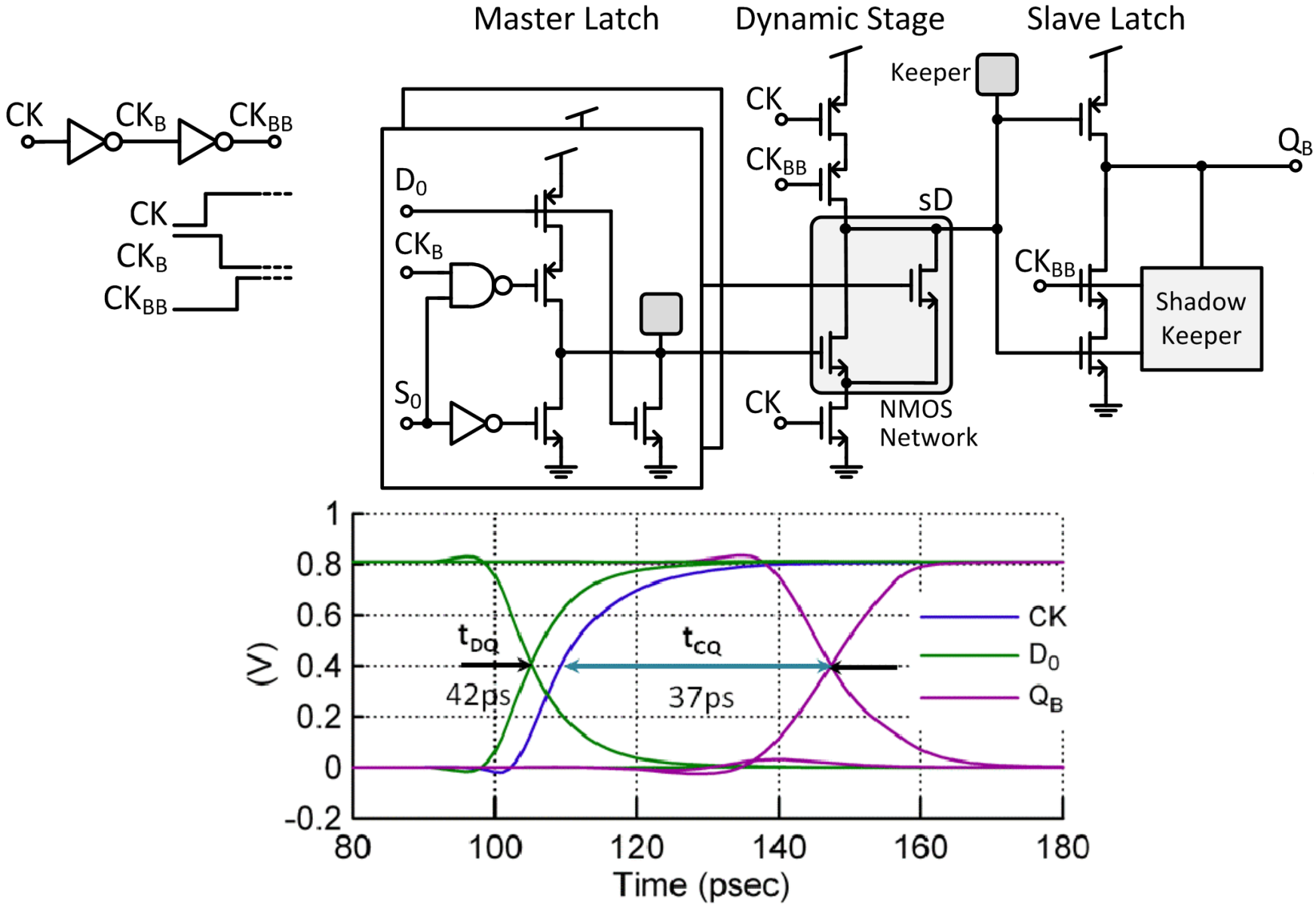
- Overview
- Methodology
- Clock and power distribution
- **Latching elements**
- Custom macros
- Results and summary

Pseudo single-phase flip-flop (PSPFF)

- ❑ Sequential elements as primary speed limiters
  - ♦ Expending up to 25% of the clock period
- ❑ PSPFF is similar to dynamic pulsed-latch
  - ♦ Superior EDP and latency *vs.* master-slave and pulsed-latch
  - ♦ Integrates complex logic
  - ♦ Tolerant to variability
    - ○ Pulse replaced with a transparent-low latch



PSPFF incorporating a 2:1 Mux



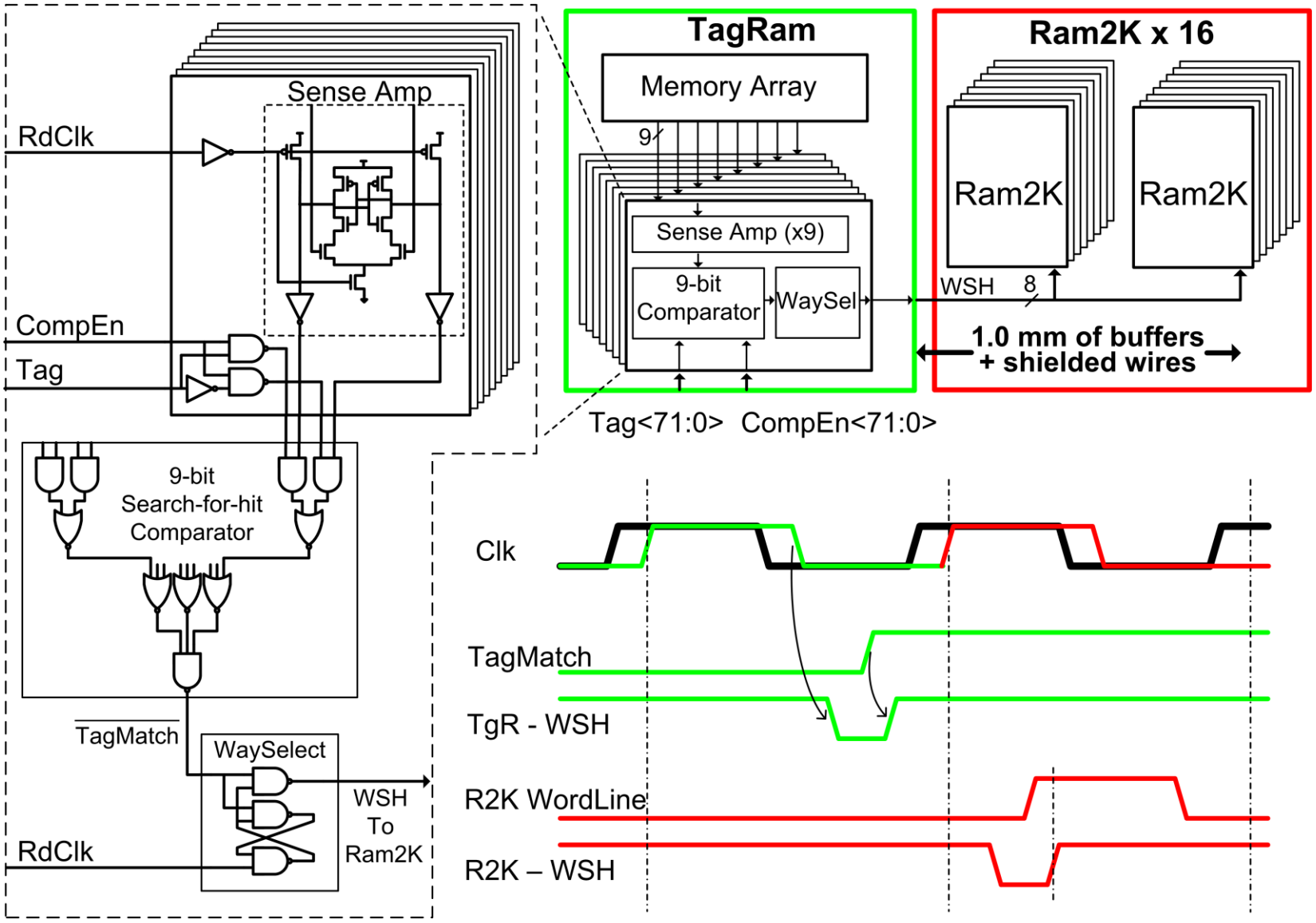
Agenda

- Overview
- Methodology
- Clock and power distribution
- Latching elements
- **Custom macros**
- Results and summary

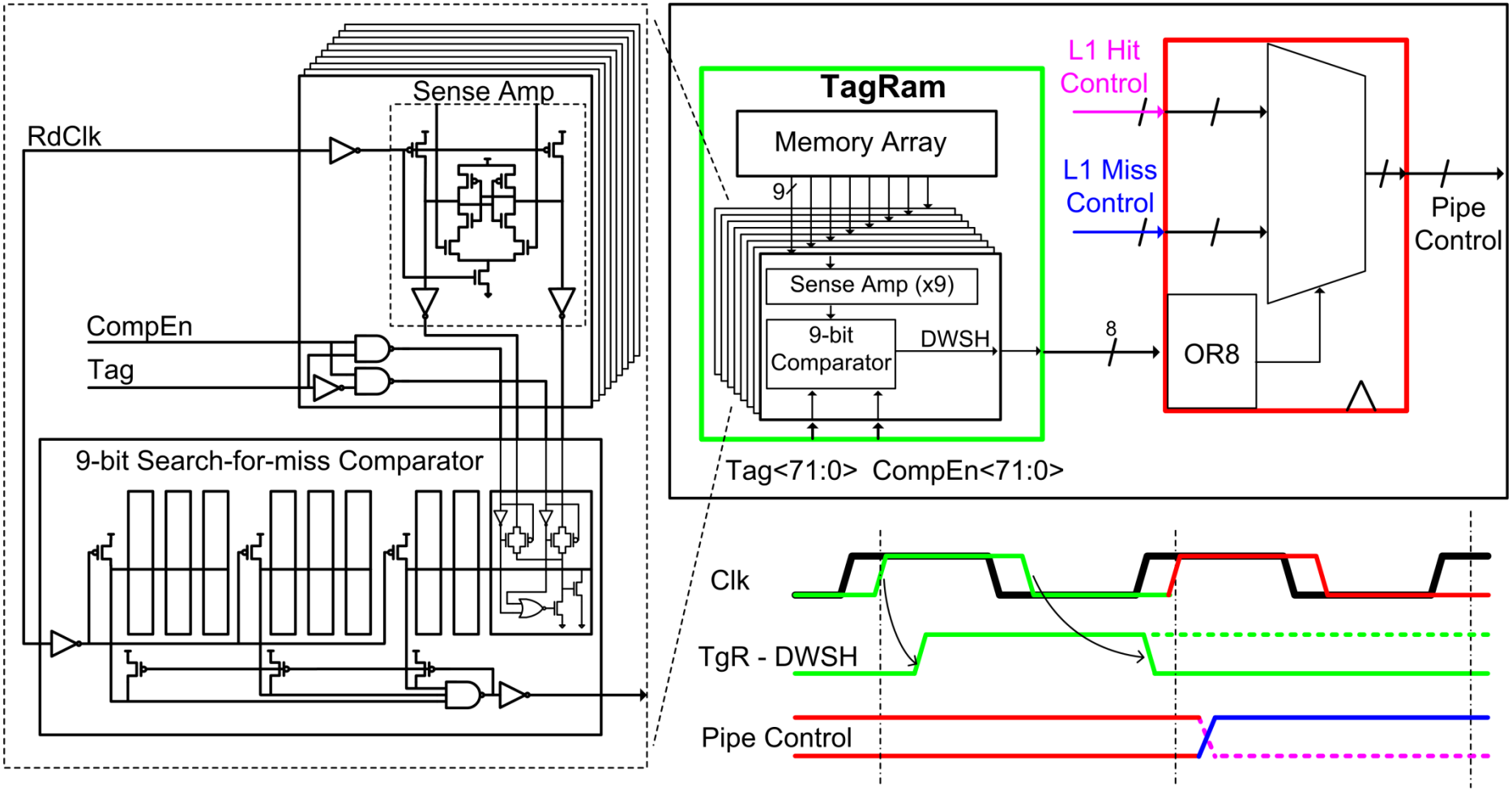
RAM2K and TAG RAM

- PMD includes a 2KB RAM (R2K) as its primary memory
  - Instantiated over 200 times in each PMD
- Handles late-arriving addresses with no added latency
  - Column selects are monotonic
- A 4 Kb Tag RAM (TgR) additionally comprises:
  - Eight sets of 9-bit comparators with two different implementations
    - One-hot (monotonic) bus as R2K column-select
    - A dynamic miss followed by an “OR8” PSPFF for hit-prediction pipeline control

Critical path - TgR to R2K

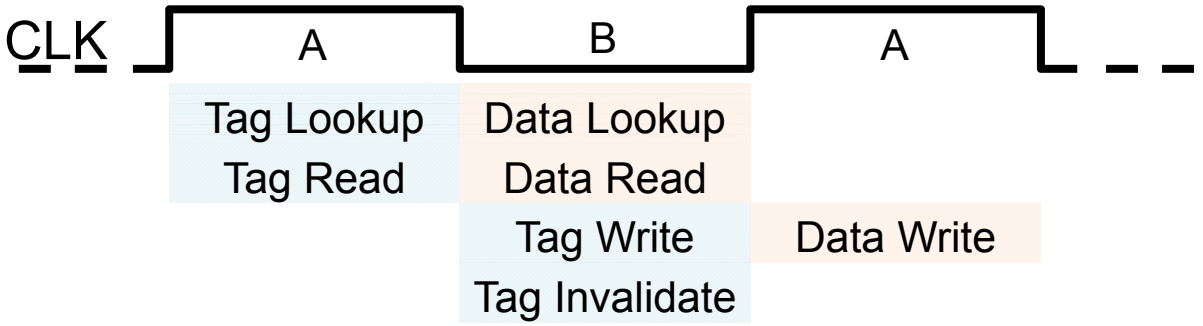


Critical path – hit prediction logic

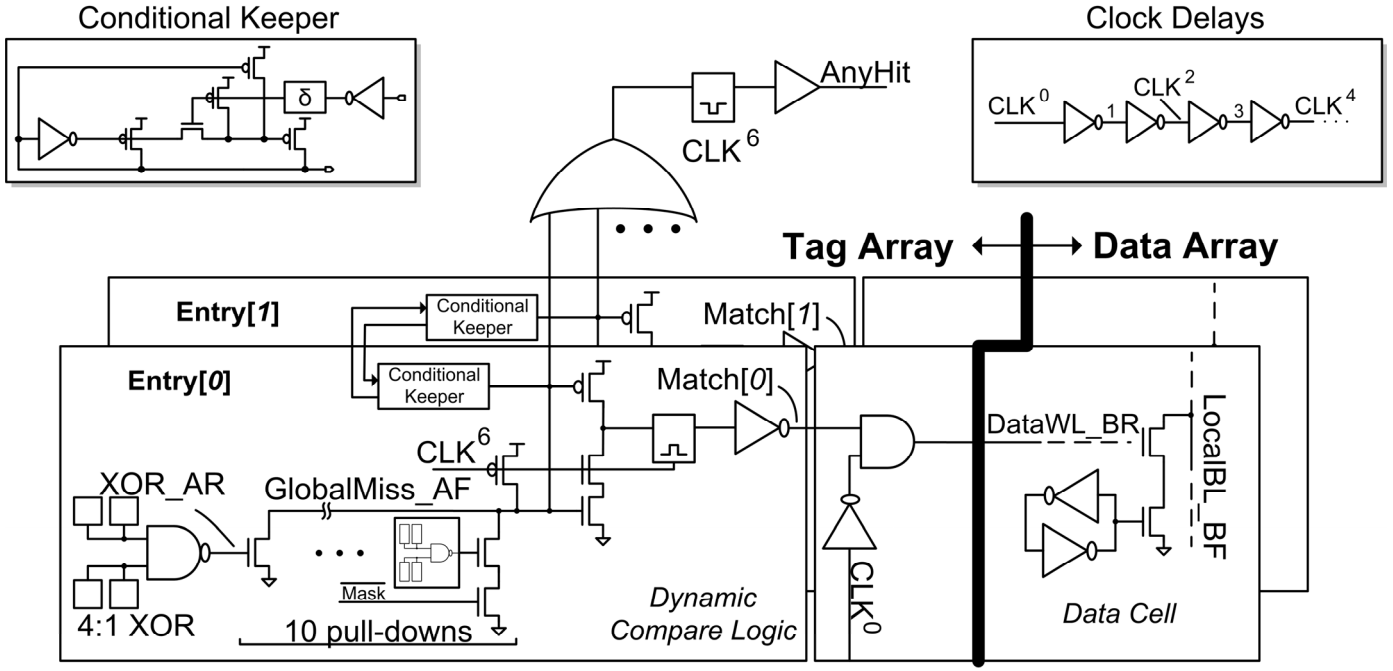


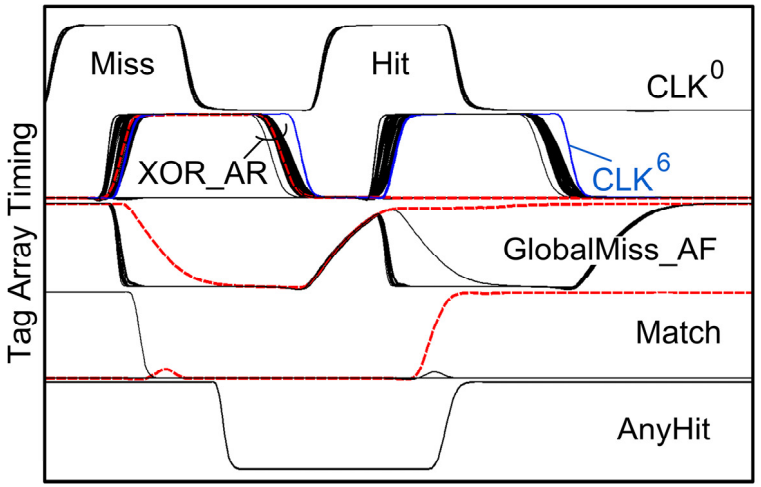
Translation Look-aside Buffer

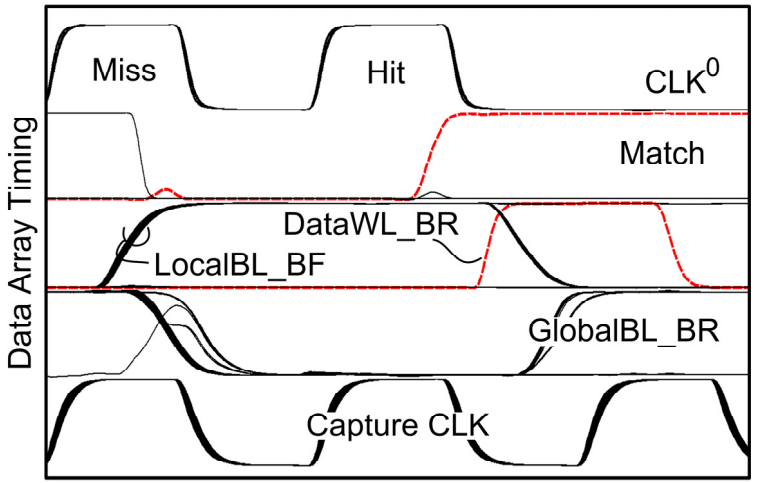
- Fully associative, single-cycle, with 20 entries
- Each entry comprises 38-bit tag and 46-bit data
  - Supports 4KB, 64KB, and 1MB dynamic page size with internal tag masking and offset forwarding
- Concurrently supports 1 write, 1 read, 2 content searches, and up to 20 invalidates
- Adopts a static cross-entry any-hit detection
- Phase-pipelined operation with pervasive use of clock-delayed domino circuits



Translation Look-aside Buffer





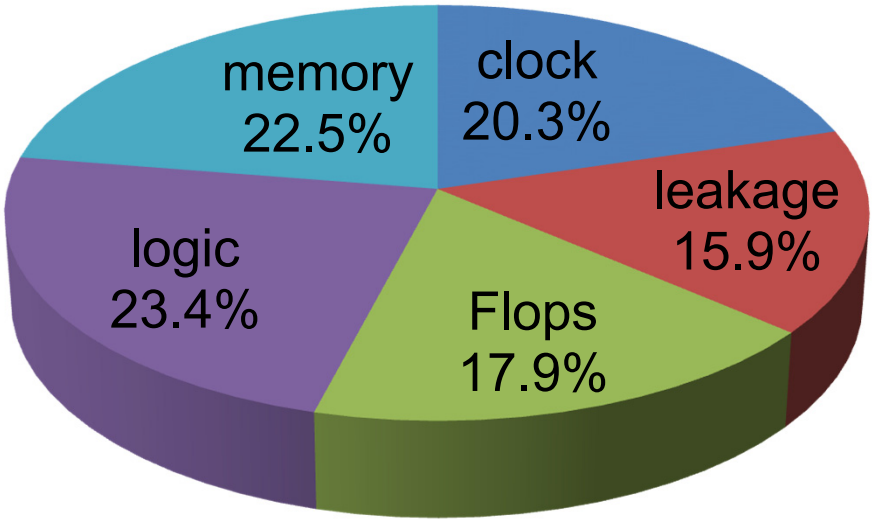


Agenda

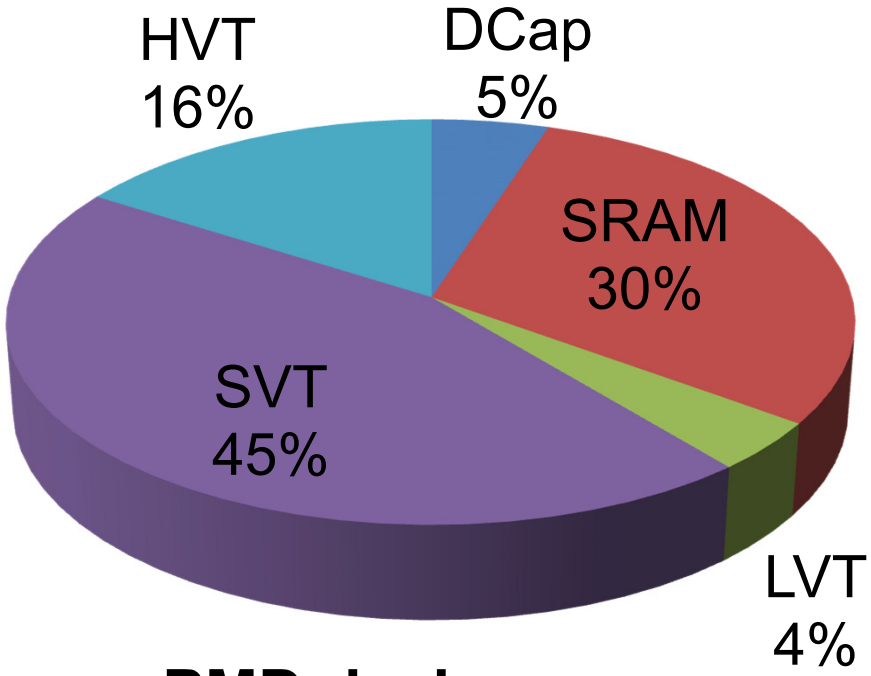
- Overview
- Methodology
- Clock and power distribution
- Latching elements
- Custom macros
- **Results and summary**

Power, area and device usage

Metric	Value
Transistor Count	84 M
Area	14.8mm^2
Power	4.5 W

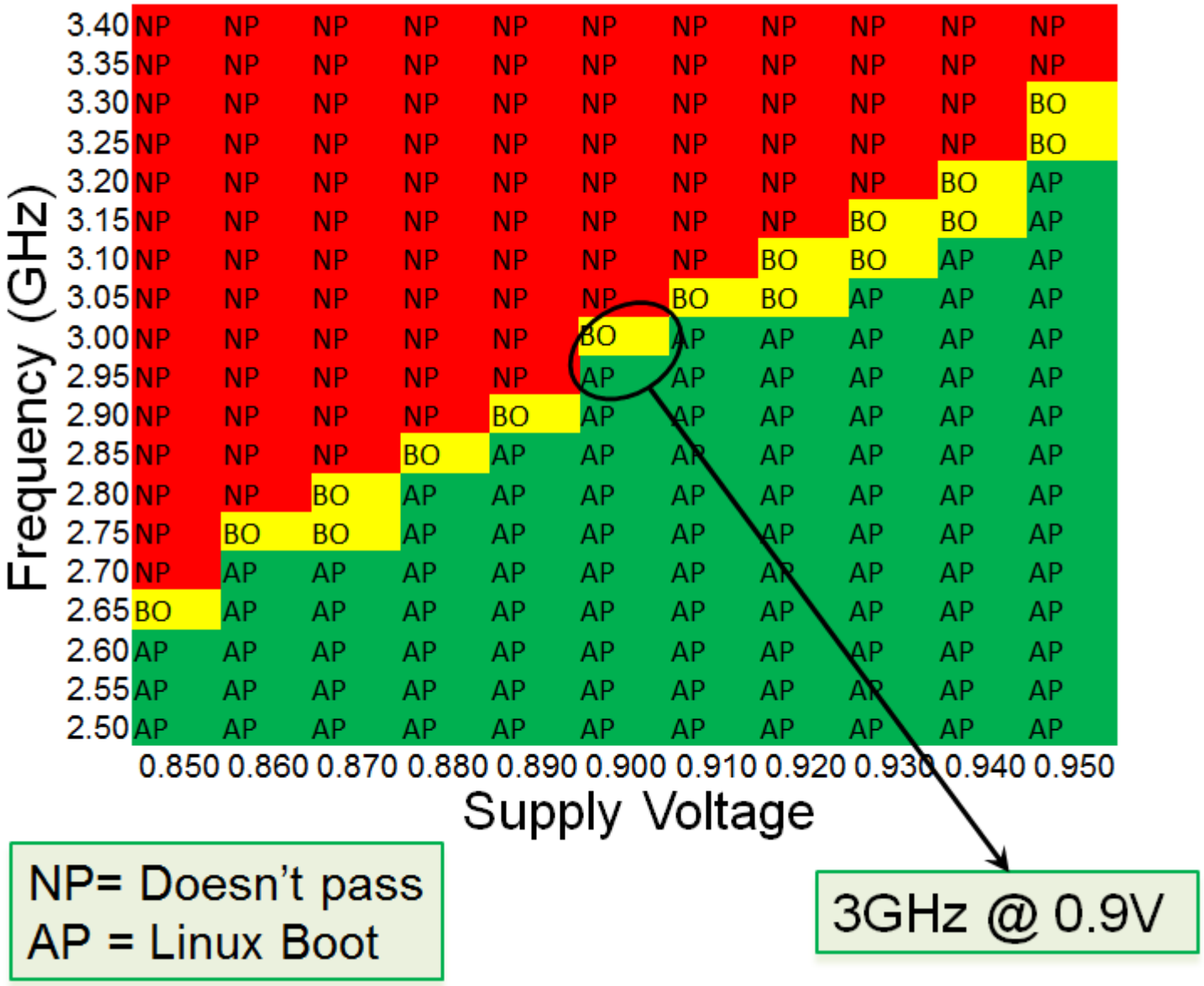


PMD power breakdown



PMD device usage

Shmoo plot



NP= Doesn't pass
AP = Linux Boot

3GHz @ 0.9V

Summary

- Design methodology and circuit techniques for the first server-class 64-bit 8-core ARM processor were presented
- Implemented in a 40nm bulk CMOS technology, the processor complex operates at 3GHz and nominally consumes 4.5W per PMD

Haswell: A Family of IA 22nm Processors

Nasser Kurd, M. Chowdhury, E. Burton, T. Thomas, C. Mozak, B. Boswell, M. Lal, A. Deval, J. Douglas, M. Elassal, A. Nalamalpu, T. Wilson, M. Merten, S. Chennupaty, W. Gomes, R. Kumar

Intel, Hillsboro, OR

nasser.a.kurd@intel.com

Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

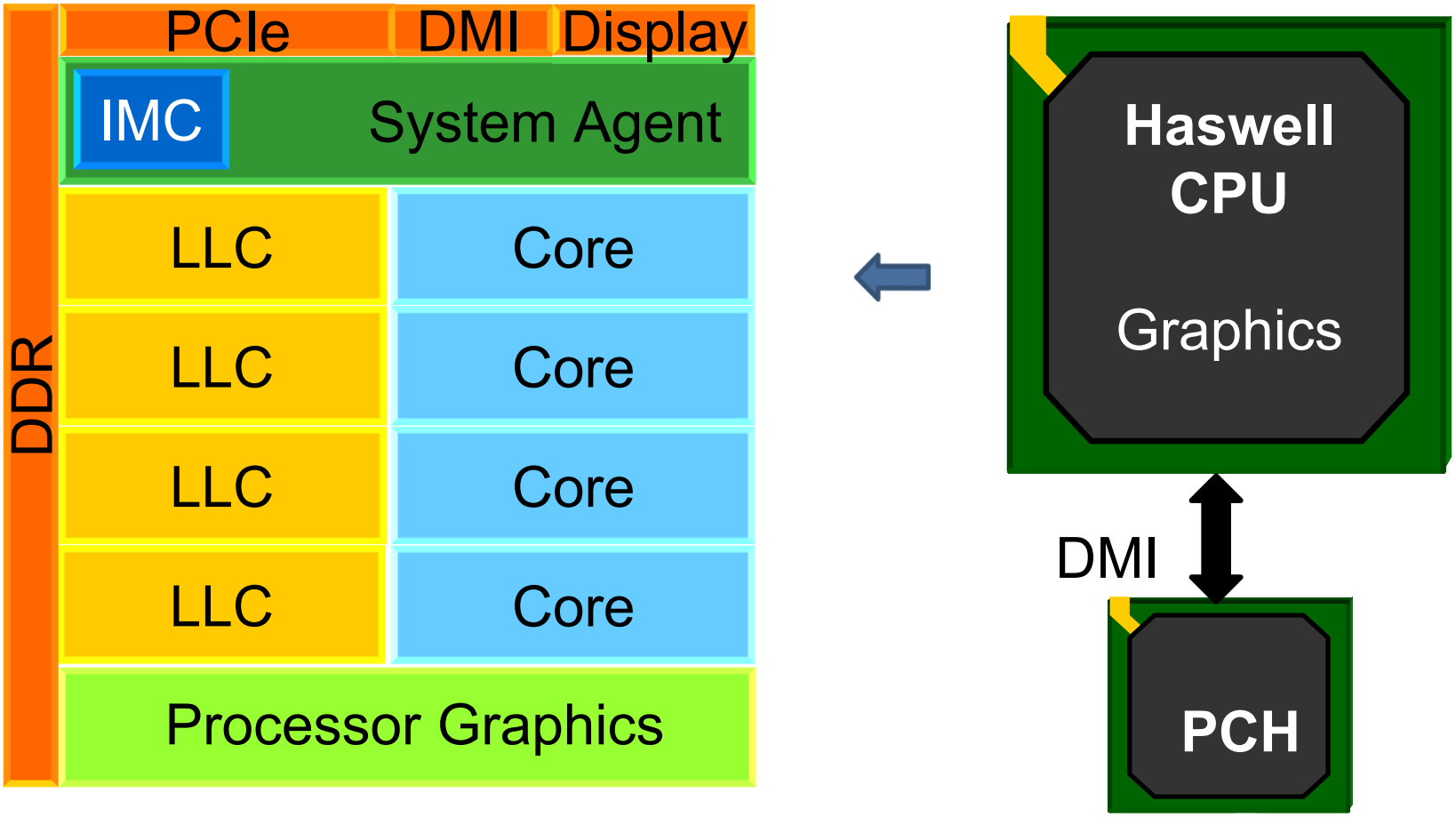
Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

The Haswell Overview

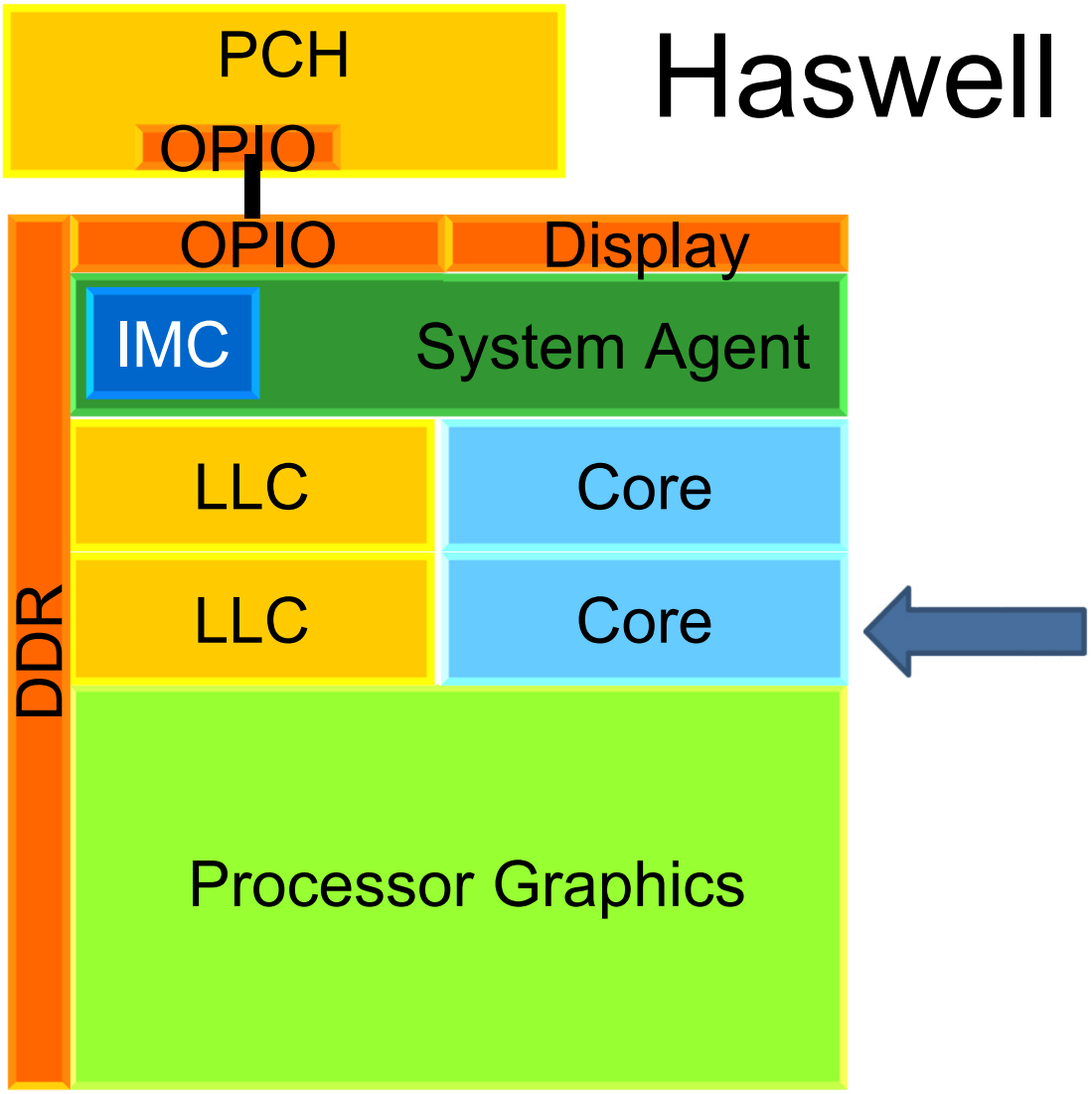
- New Architecture for 4th Generation Intel Core Processor Family
  - 22nm Intel® tri-gate process technology
- Haswell vision
  - Platform integration
  - Lower power
  - New scalable form factors:
    - Fan-less to Desktop

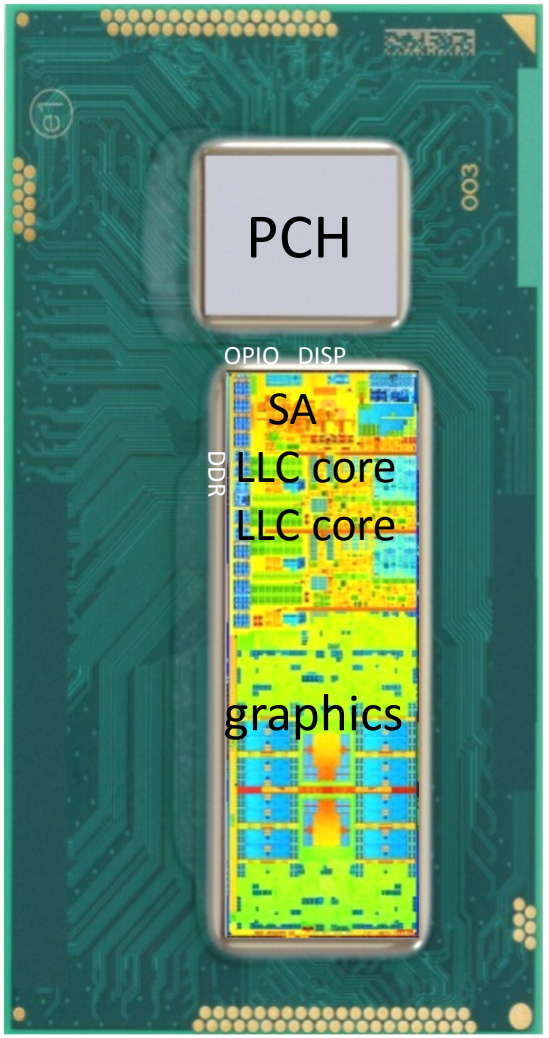
Haswell + PCH (2-Chip)



- X16 PCIe gen2/gen3, 2Ch. DDR3,HDMI 1.4, DP 1.2, eDP
- Fully Integrated Voltage Regulator (FIVR)
- **Improved** graphics, cores (2-4)

Haswell Ultrabook®

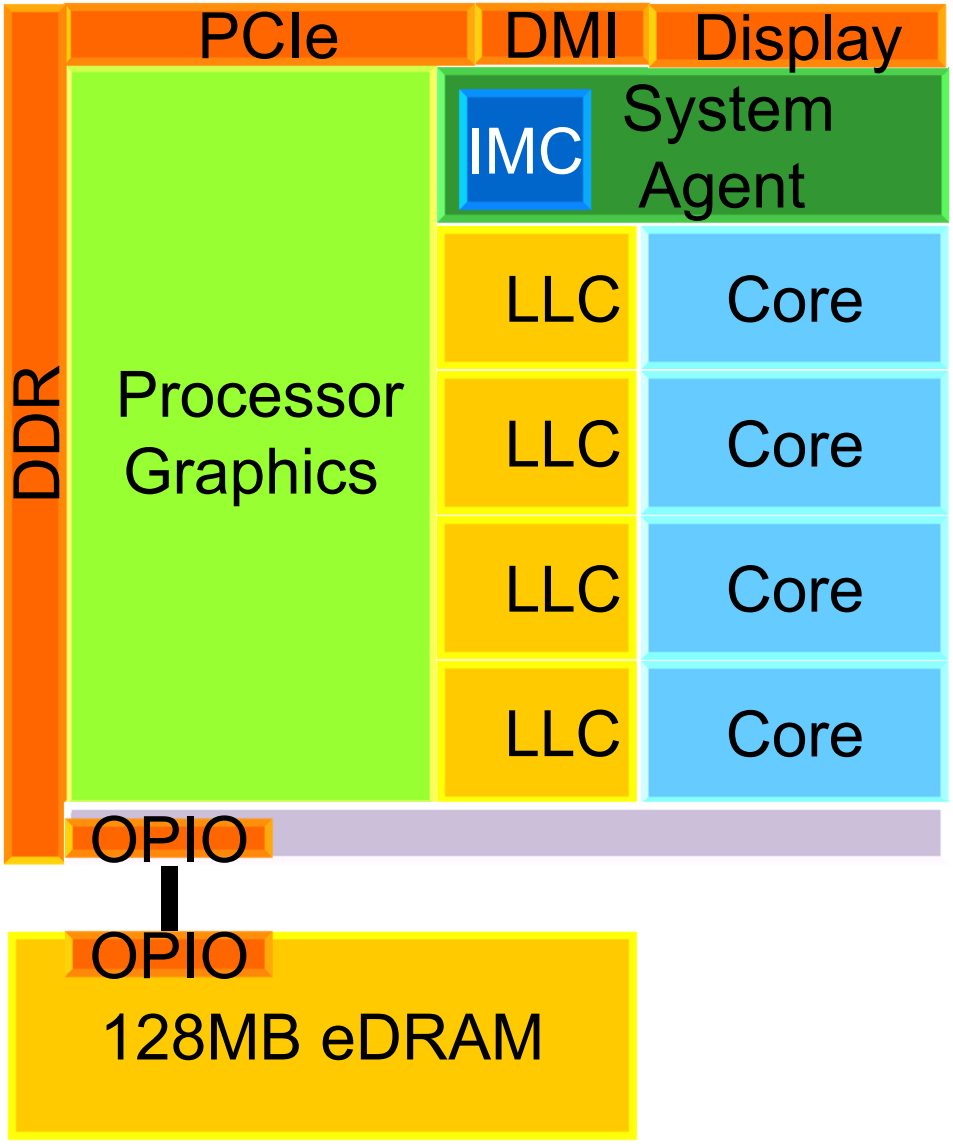


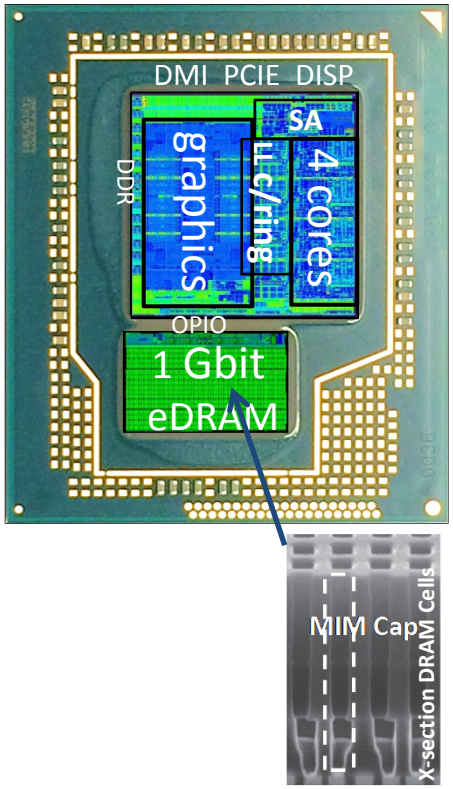


- Integrated: Low power PCH, FIVR
- 2 CH LPDDR, On Package I/O (OPIO)
- Lower power states

*Graphics and Cores numbers vary

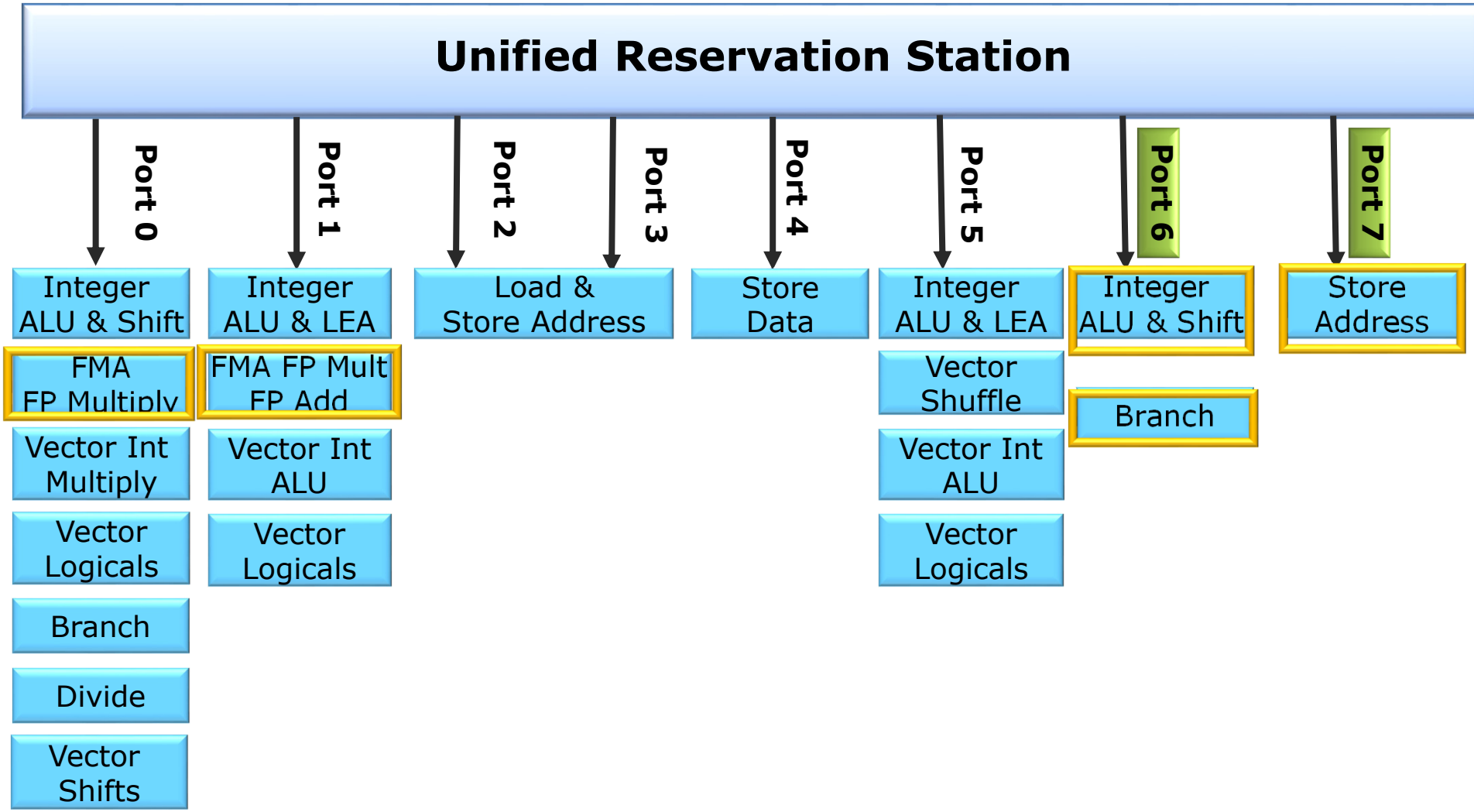
Intel® Iris™ Pro graphics





- Integrated eDRAM (128MB), FIVR
- On Package I/O (OPIO)

Core Microarchitecture Enhancements



New Compute Instructions

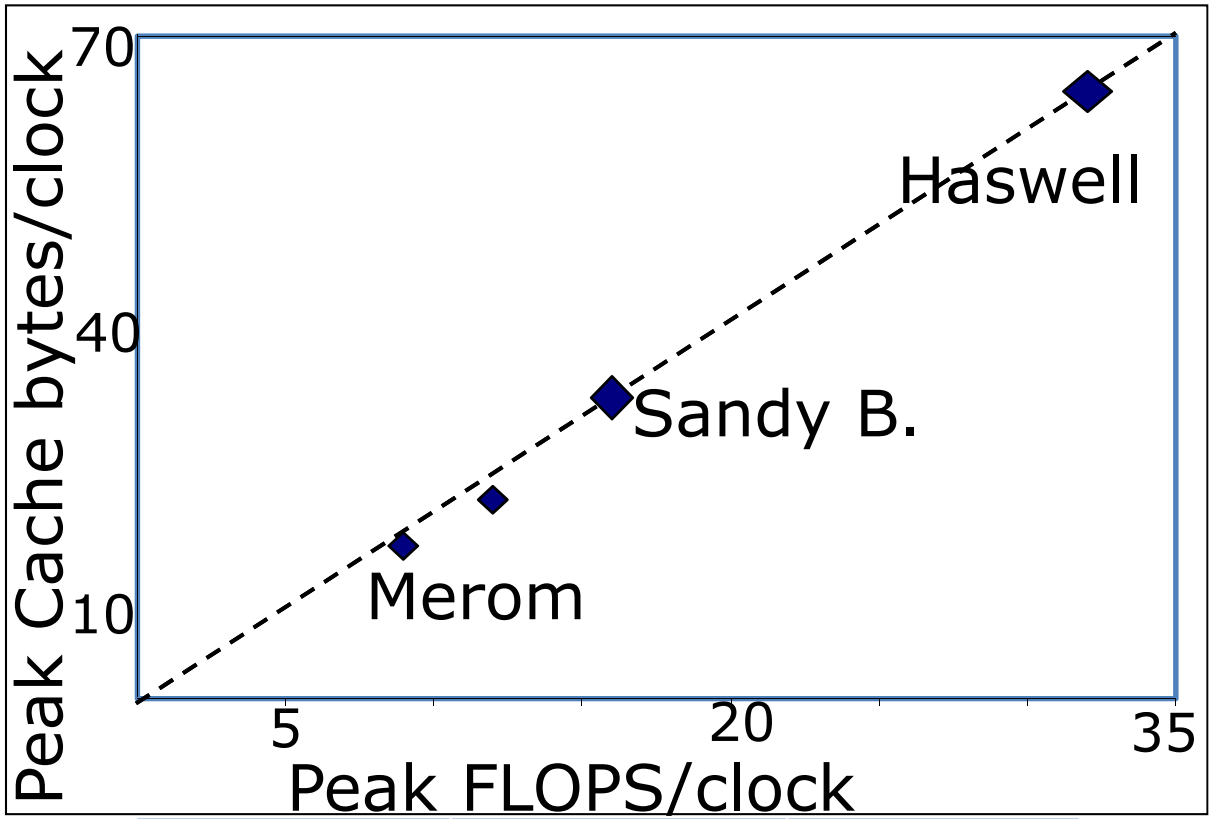
CPU Family	Instruction Set	Flops/ Core SP	Flops/ Core DP	L1 Cache BW (bytes/clk)	L2 Cache BW (bytes/clk)
Nehalem	SSE (128b)	8	4	32 (16R, 16W)	32
SandyB.	AVX (256b)	16	8	48 (32R, 16W)	32
Haswell	AVX2 (256b)	32	16	96 (64R, 32W)	64

- Intel® AVX2
  - 256-bit integer vectors
  - FMA, Full-width Element permutes, Gather
  - New Integer Instructions

Full Instruction Specification Available at http://software.intel.com/en-us/avx/

FMA & Peak FLOPS

- 2 new FMA units provide 2x peak FLOPs/cycle of previous generation
- 2X cache bandwidth to feed wide vector units
  - 32-byte load/store for L1
  - 2x L2 bandwidth
- 5-cycle FMA latency same as an FP multiply

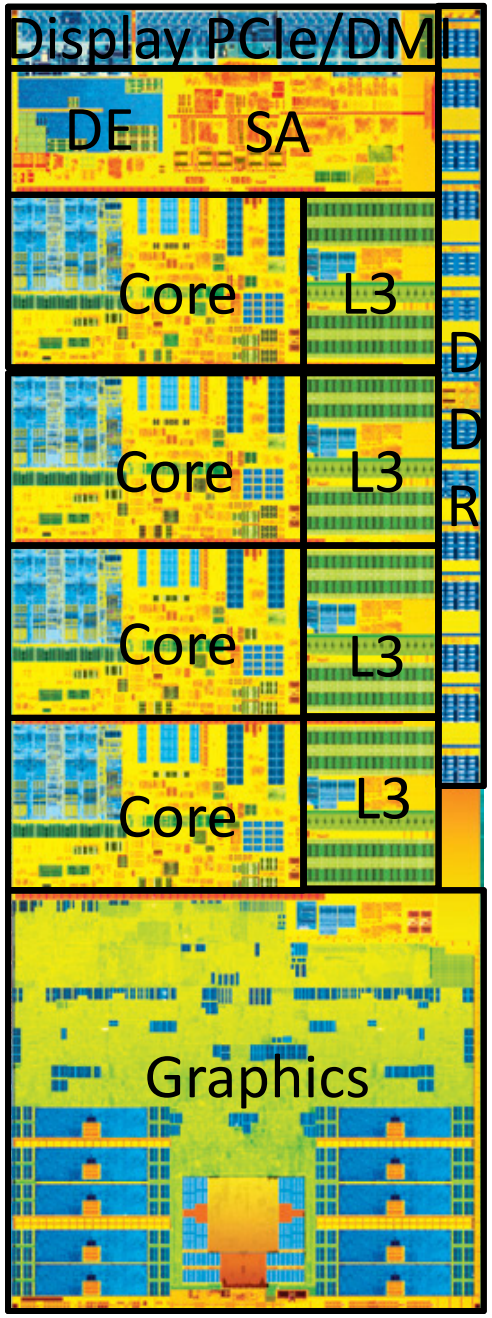


Latency (clks)	Prior gen	Haswell
MulPS, PD	5	5
AddPS, PD	3	3
Mul+Add/ FMA	8	5

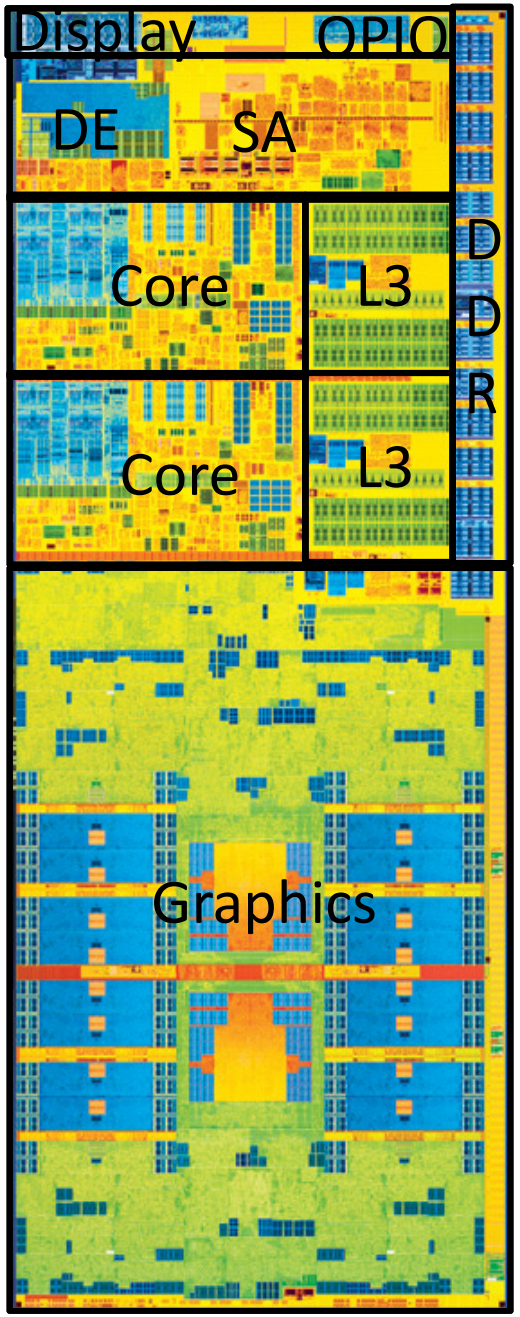
Die Photos

Technology	22nm tri-gate CMOS
Cores	2-4
Graphics	GT2-GT3
Transistor count	0.96-1.7B
*Chip size	130-260mm2
Voltages	0.7-1.1V
Nominal Frequency	1.1-3.8GHz

* Depending on core and graphics



Quad-Core



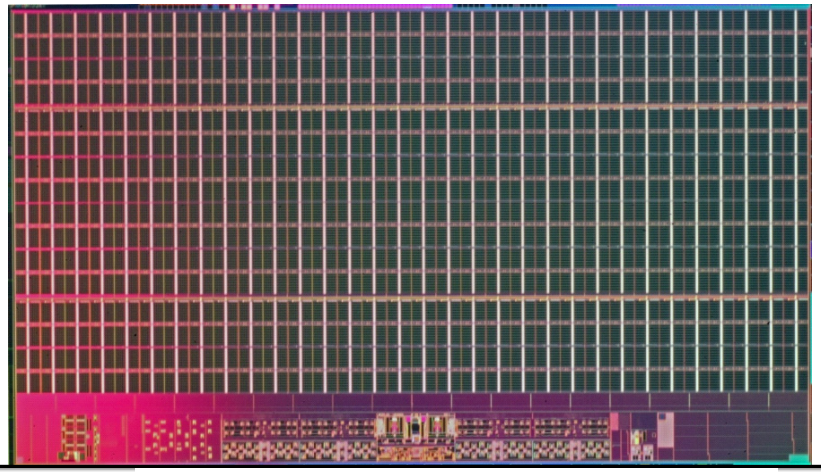
Dual-Core

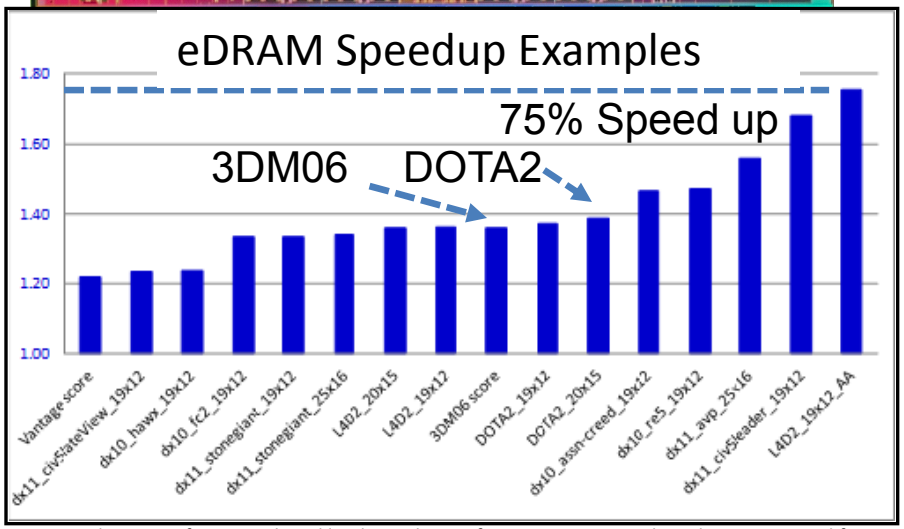
Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

eDRAM

Technology	22nmTri-Gate CMOS
Cell Size	0.029 um2
Macro Area	17.5Mb/mm2@128Mbit
Chip Org.	8 Macros(16MB) Chip:128MB array
Subarray Config.	256 wl x 1024 bit-line
Chip Size	77mm2
Supply	1 V
Clock, R. Cycle T.	1.6GHz, 3.75ns
Retention Time	100us @93C





- Extend cache hierarchy
- The CPU and eDRAM are in MCP-full-duplex on package IO

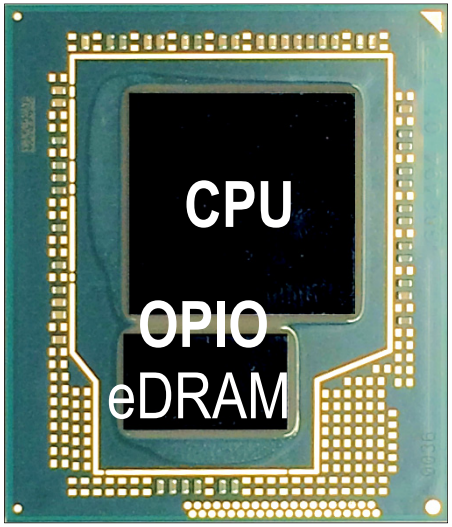
Details in paper 13.1: A 1Gb 2GHz Embedded DRAM in 22nm Tri-Gate CMOS Technology

Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

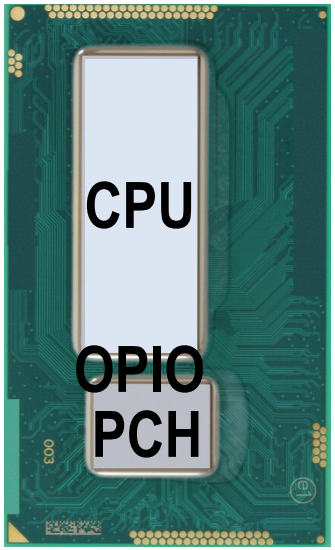
On Package I/O (OPIO)

CPU-eDRAM (6.4GT/s)



102.4 GB/s using 1W (1.22pJ/b)

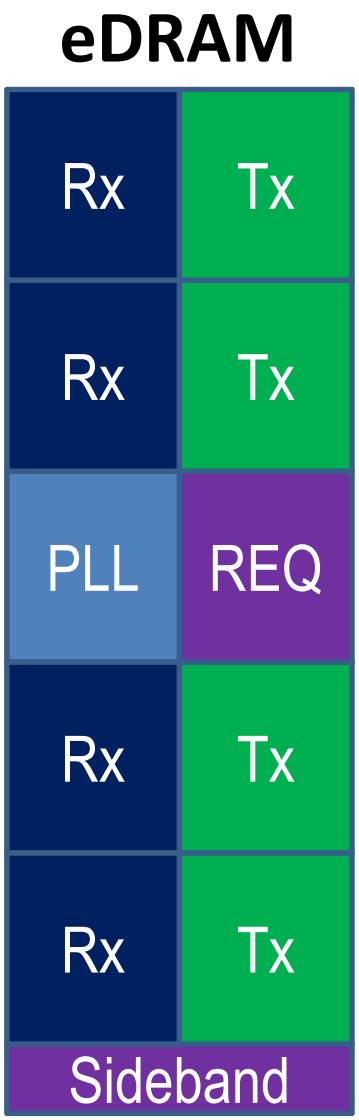
CPU-PCH (2GT/s)



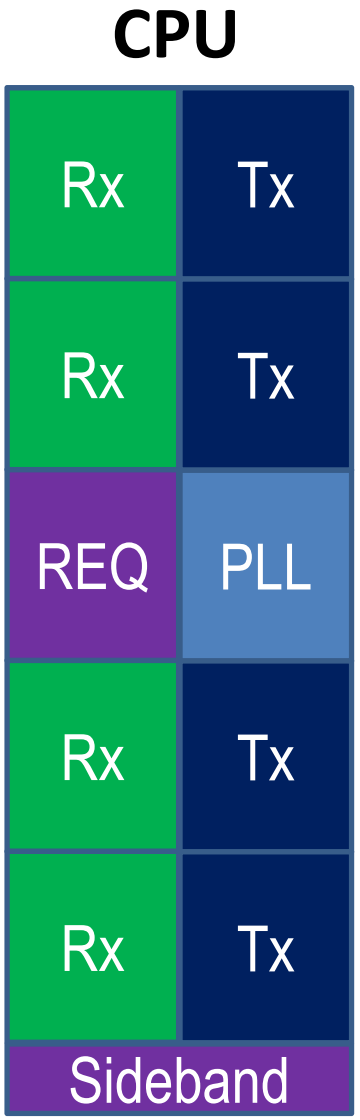
4 GB/s using 32mW (1pJ/b)

- OPIO is used for Multi-Chip Packages
  - CPU-eDRAM and CPU-PCH MCPs
  - MCP provides product and process flexibility
  - Reduces platform power and form factor
- OPIO is a single-ended high bandwidth I/O
  - Very low power with close chip assembly
  - Simplified CMOS I/O and clocking circuits
  - No Rx termination, reduced ESD protection

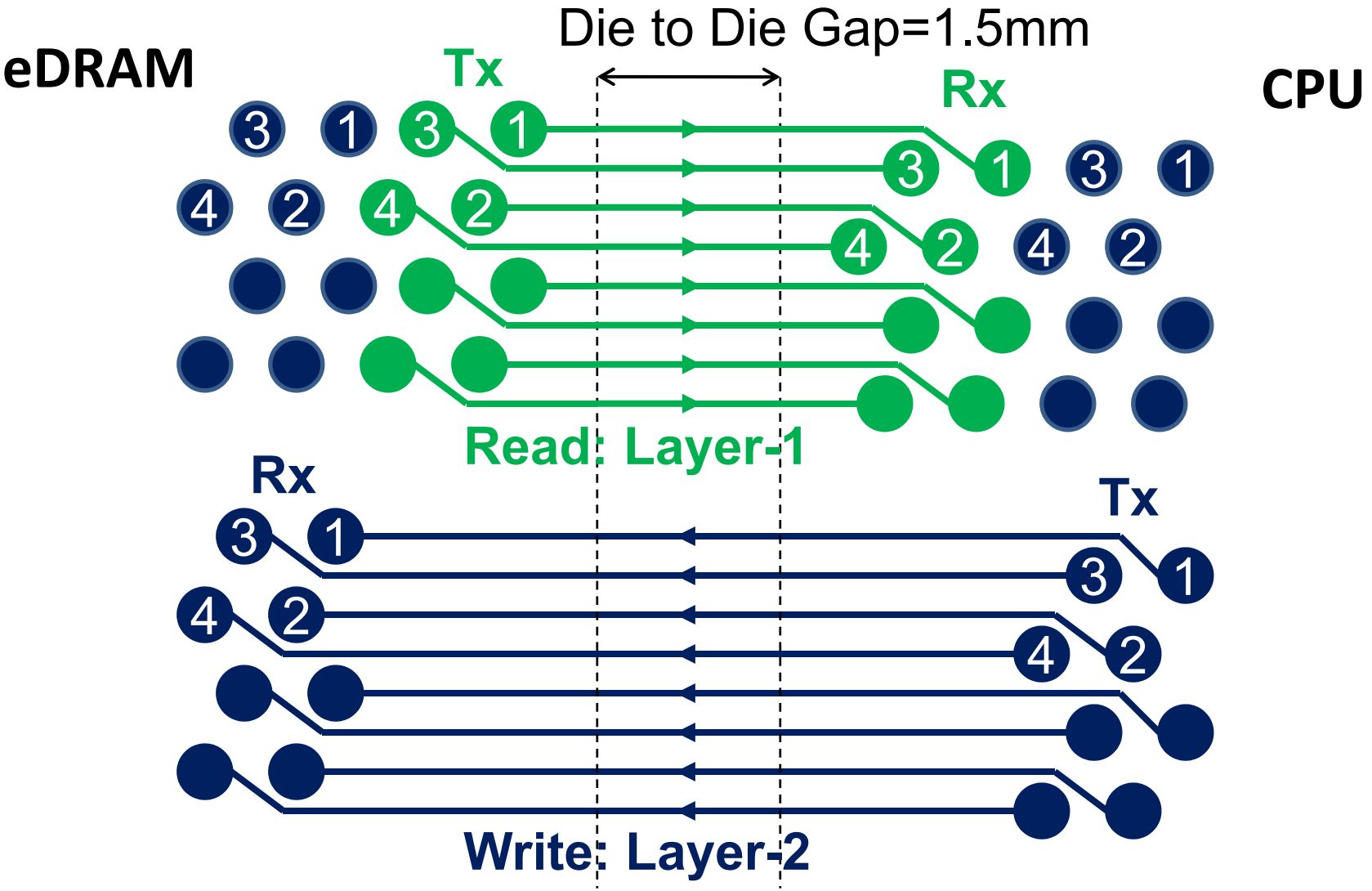
CPU-eDRAM OPIO



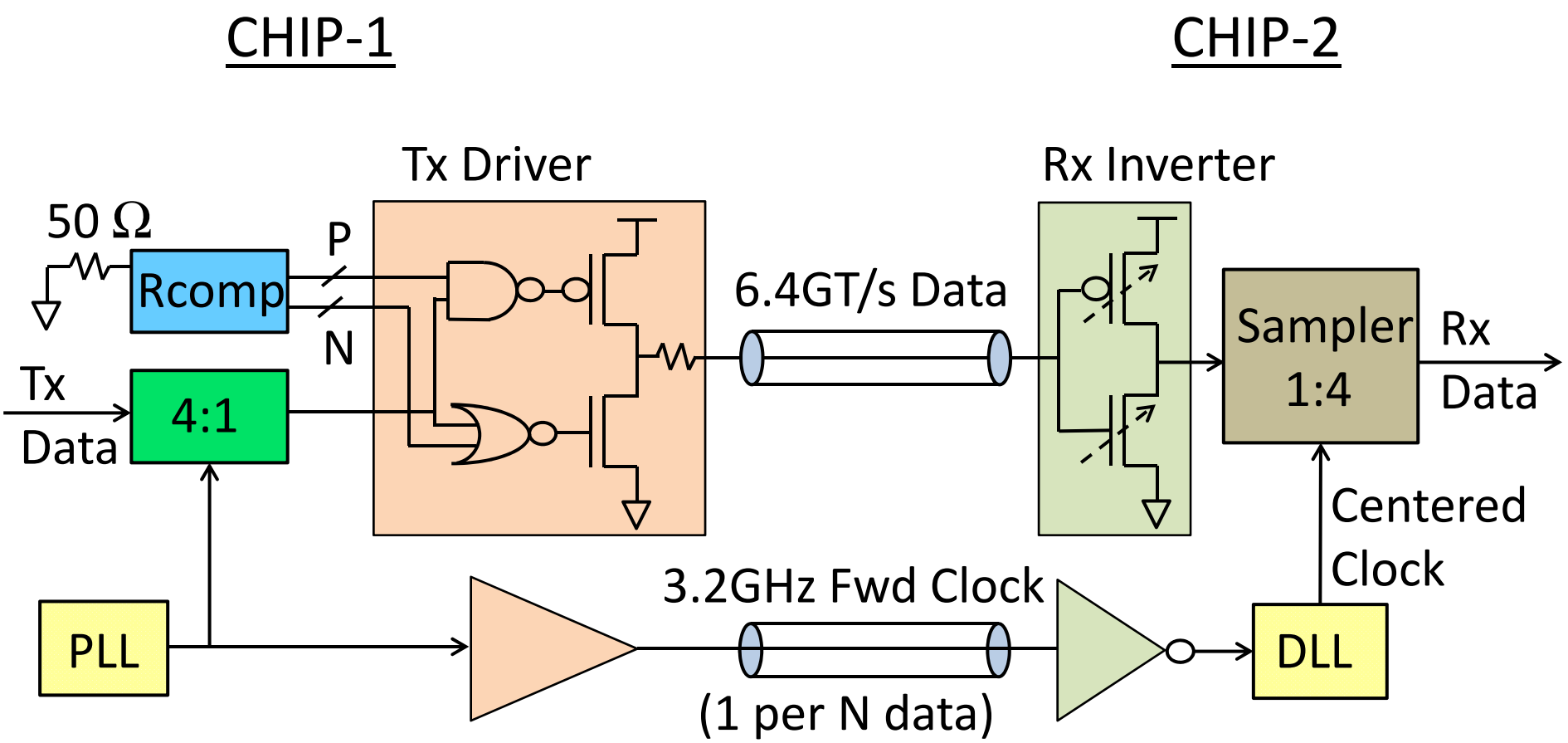
- 8 data clusters, each with
  - 16 data bits
  - Fwd clock, ECC, Valid
- 1 Request cluster
  - Command & Address
- PLL block & sideband
- 64 bit data each way
  - at 6.4GT/s
- 2 x 8 x 6.4 = 102.4 GB/s



OPIO Length Matched Routing



Simplified OPIO PHY

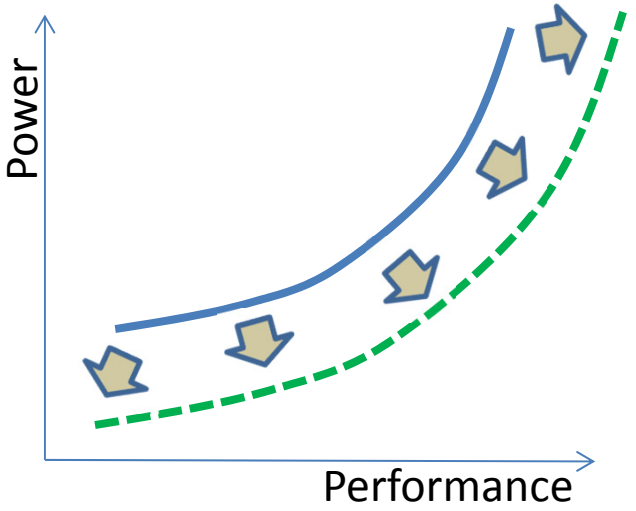


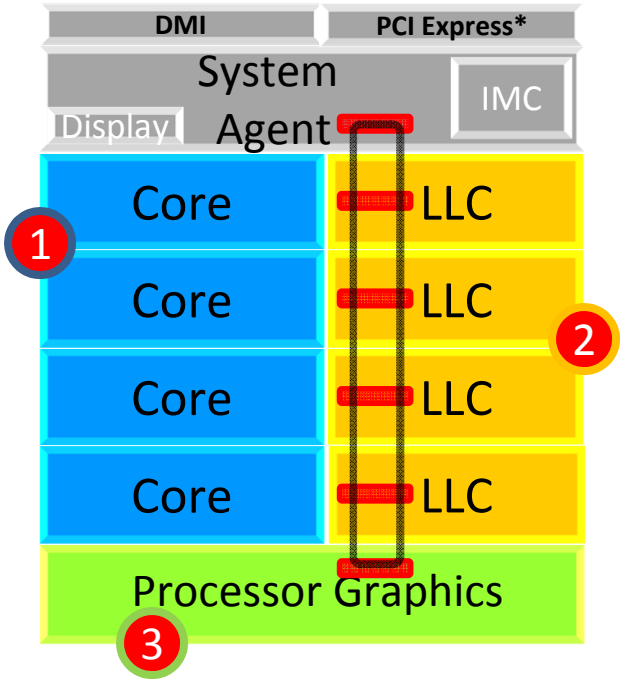
Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

Maximizing Power-Limited Performance

- Extended operating range
  - Power efficient features: better than voltage / frequency scaling
  - Continued focus on gating unused logic and low-power modes
- Independent frequency domains
  - Cores separated from LLC+Ring for fine-grained control
  - Power Control Unit dynamically allocates budget when power-limited
  - Prioritization based on run-time characteristics selects domain with the highest performance return



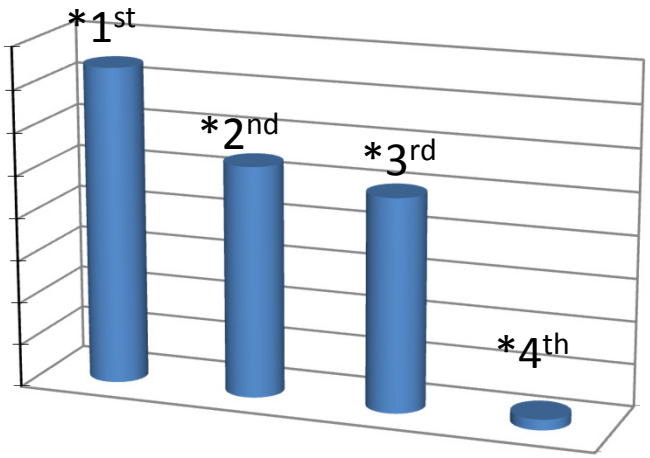


Maximizing Battery Life

- Deeper idle states, lower active power
  - Focus on turning off blocks that are not required. Example: C7
    - All clocks stopped, voltage removed from the majority of the CPU
    - C7 engaged even when display is active
  - Faster state transition times by ~25%
- Smarter low power states
  - New S0ix idle states with idle power enabling tablet FF with Core®
  - More C-state intelligence

CPU Idle Power

(High volume mobile CPUs)



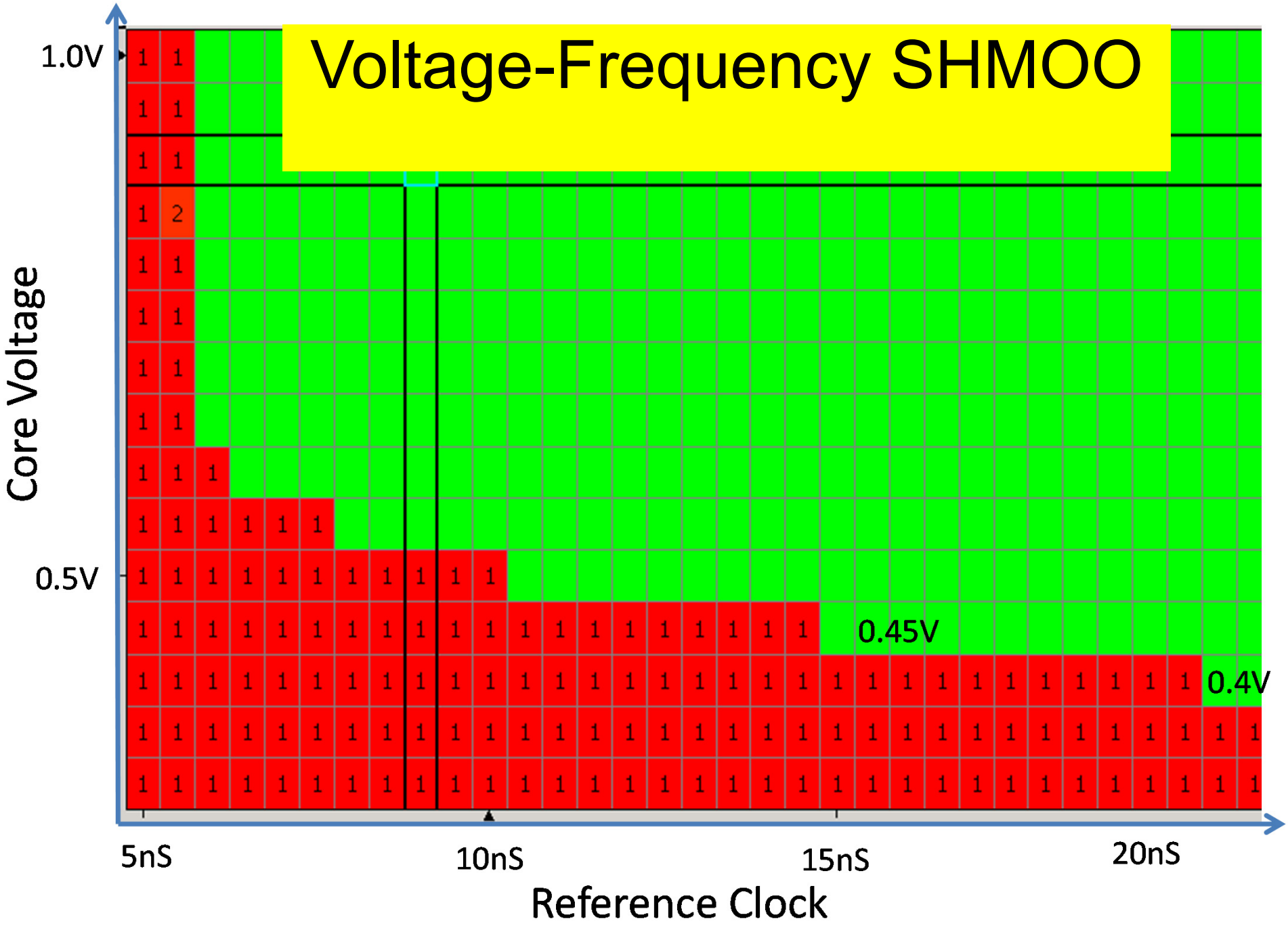
* Core generation

Average power consumption reduced by a factor of 20

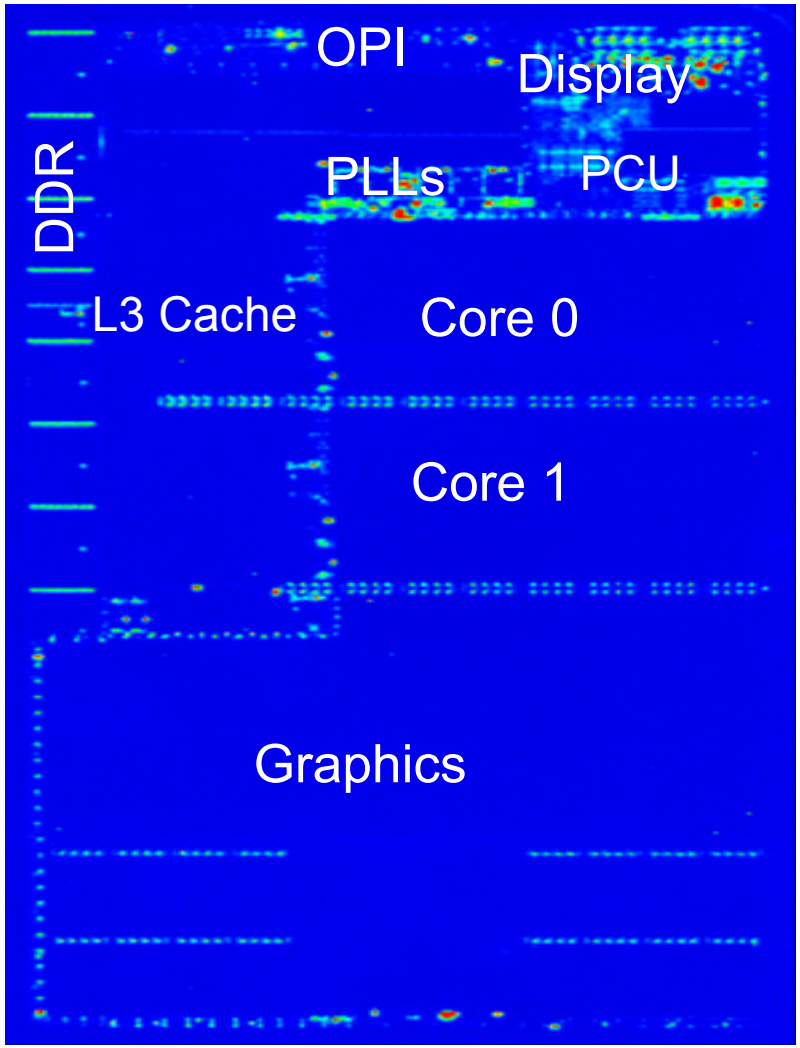
Deep Package States

State	CPU Action	**Power
C0	Core & Graphics units active executing instructions.	*15W
C6	Core's & Graphics Engine are powered off. DDR is in self refresh. Most PLL's off.	320mW
C7	System agent gated, DDR IO logic gated.	250mW
	L3 is flushed, critical arrays to be powered by the Sustain Rail. Turn off Reference clock	
C8	Display, IO rails off.. System Agent is gated. Vin is lowered to 1.2v.	77mW
C9	Vin is lowered to 0V.	18mW
C10	VR controller power optimizations	18mW

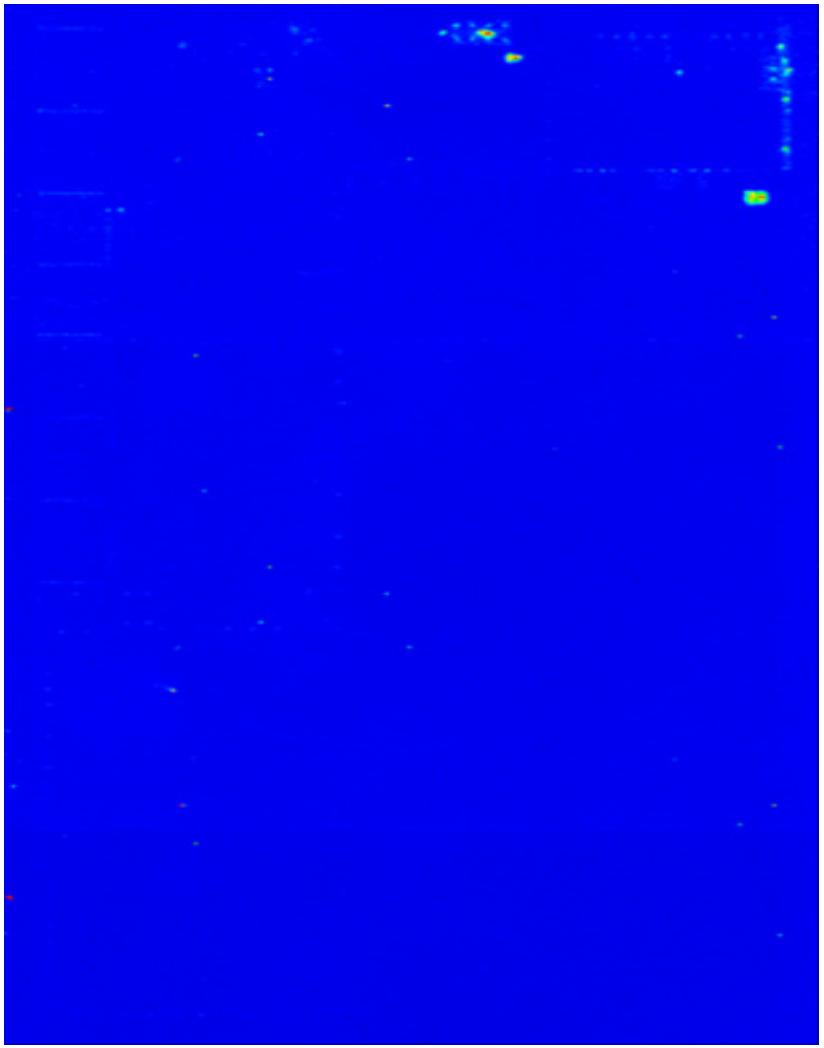
***TDP: 15W Thermal Design Power SKU ** typical**



IREM Images of Idle States



C7

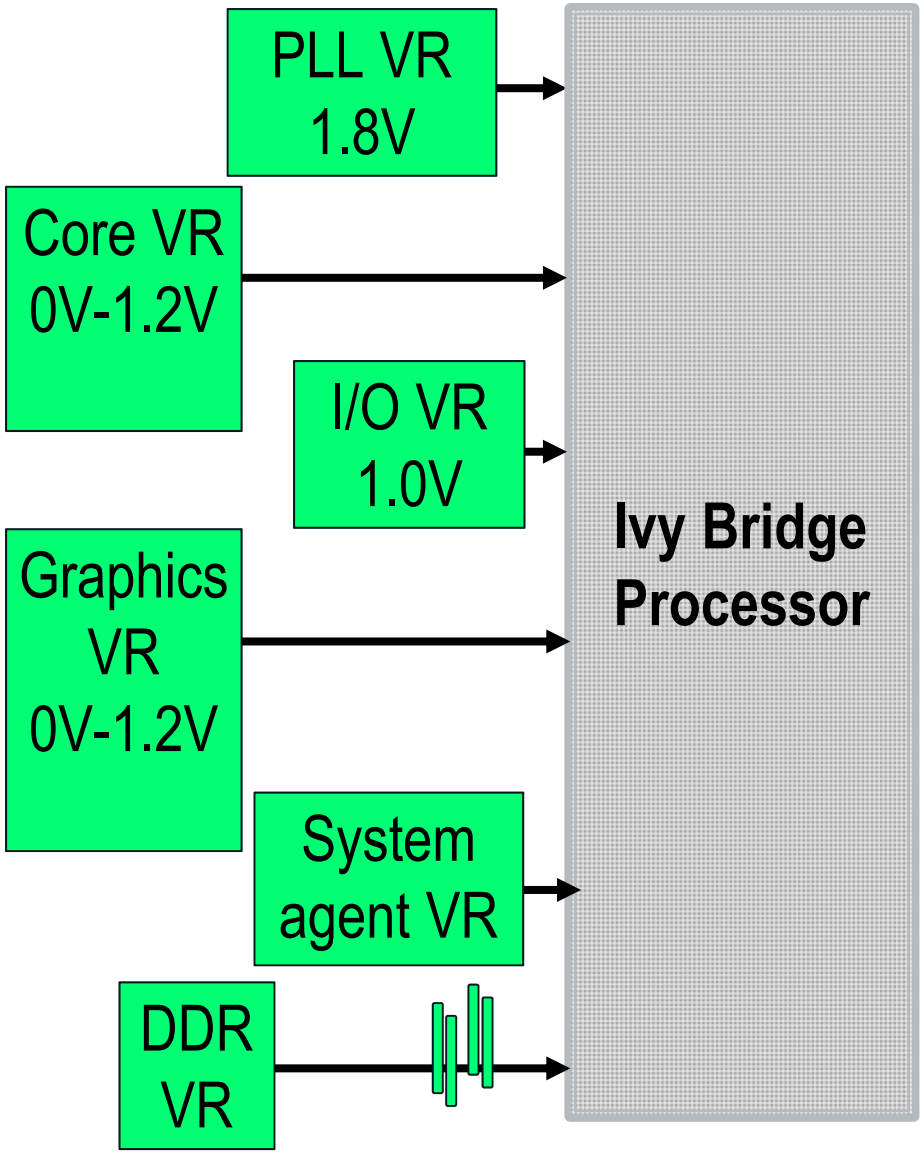


C9

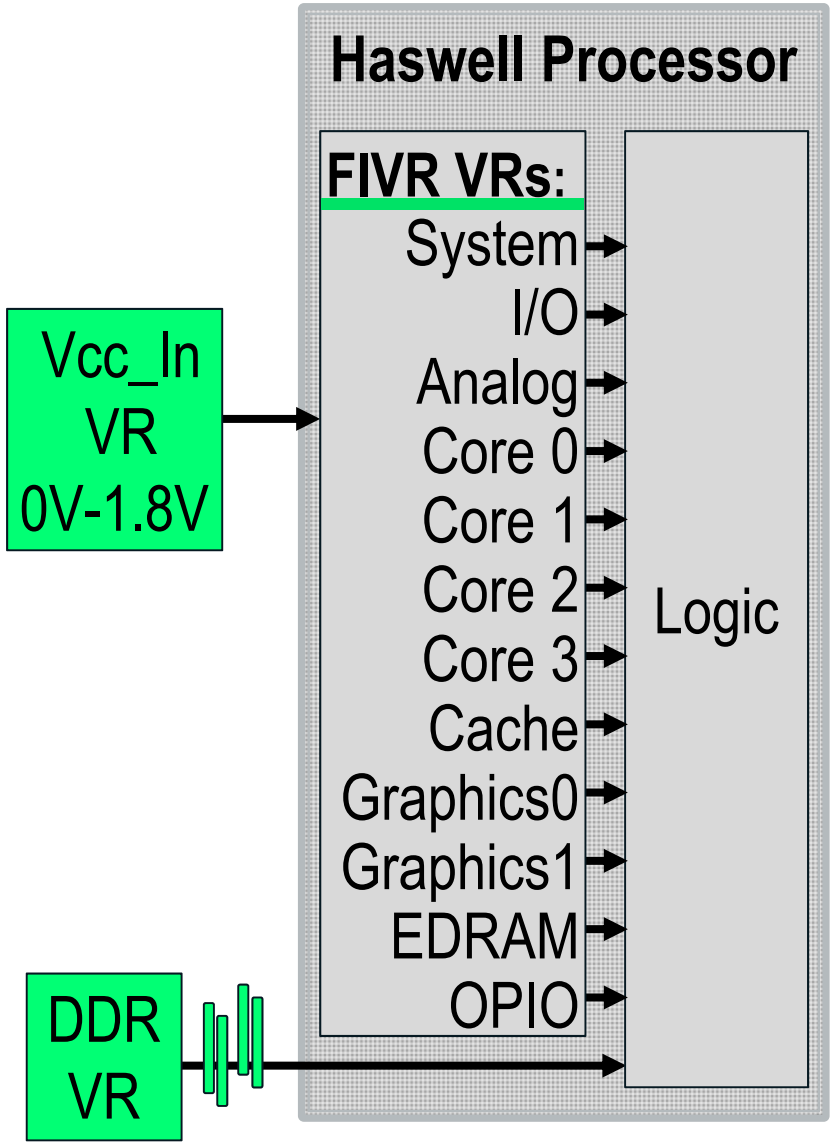
Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

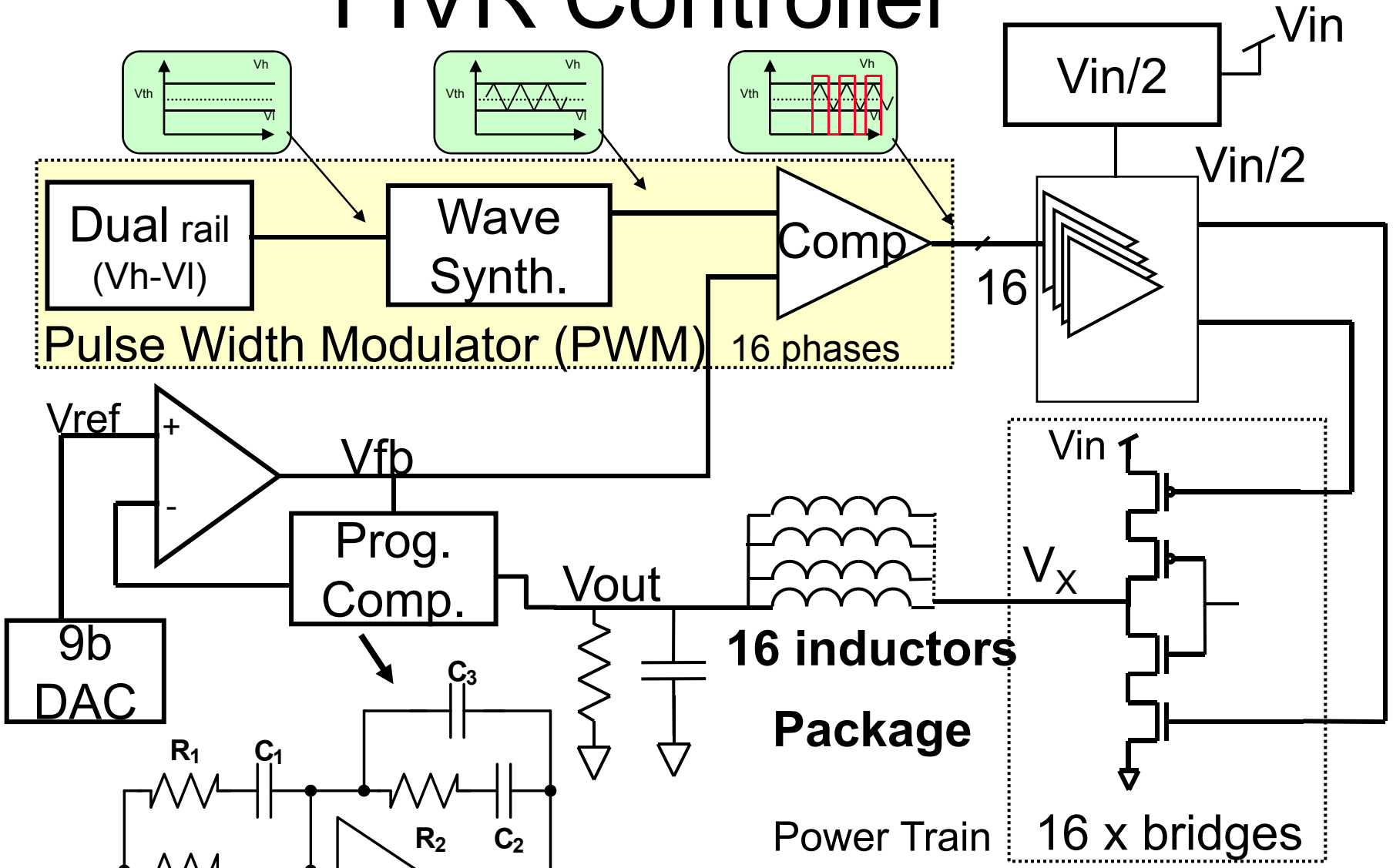
Ivy Bridge Platform



Haswell Platform

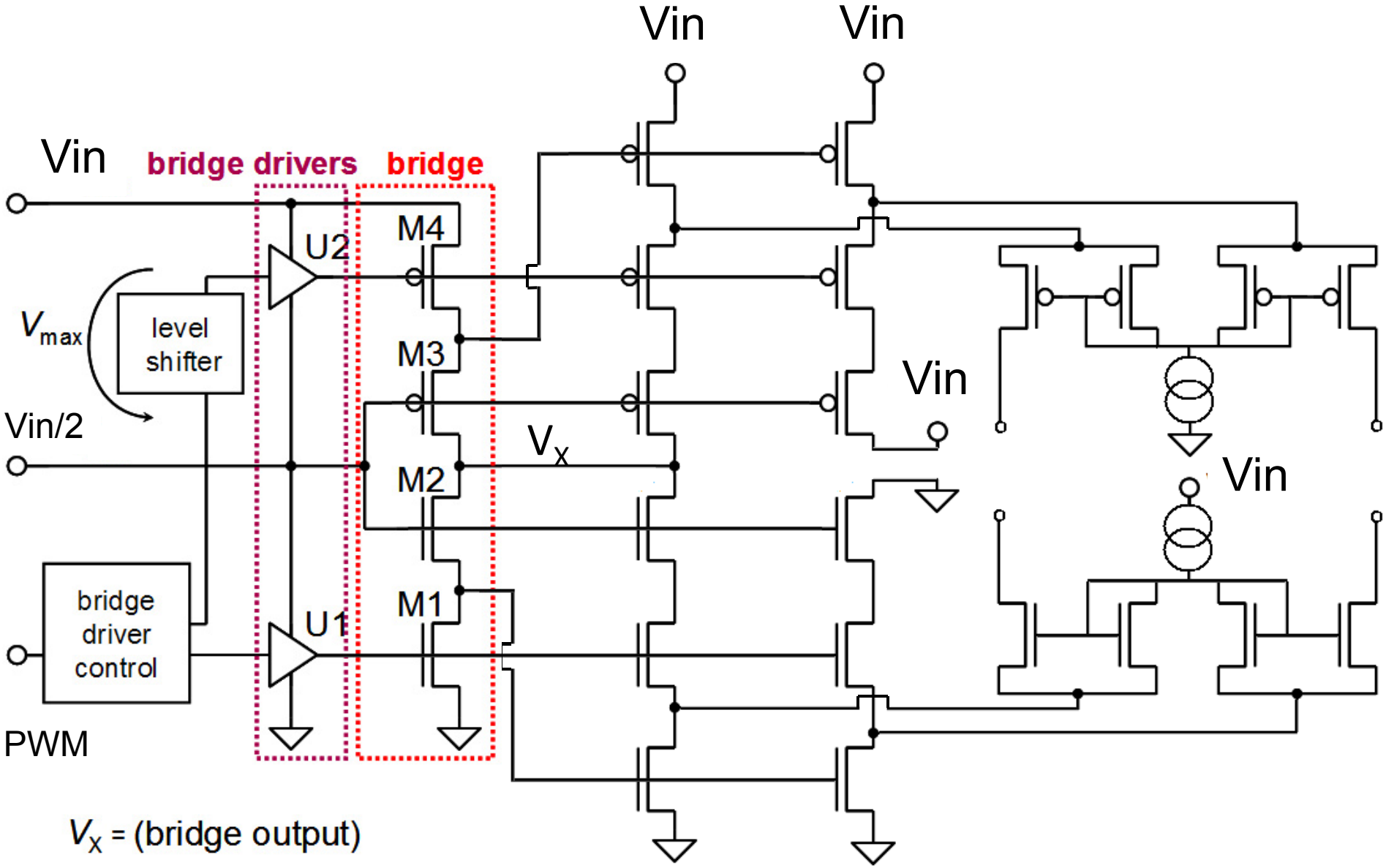


FIVR Controller

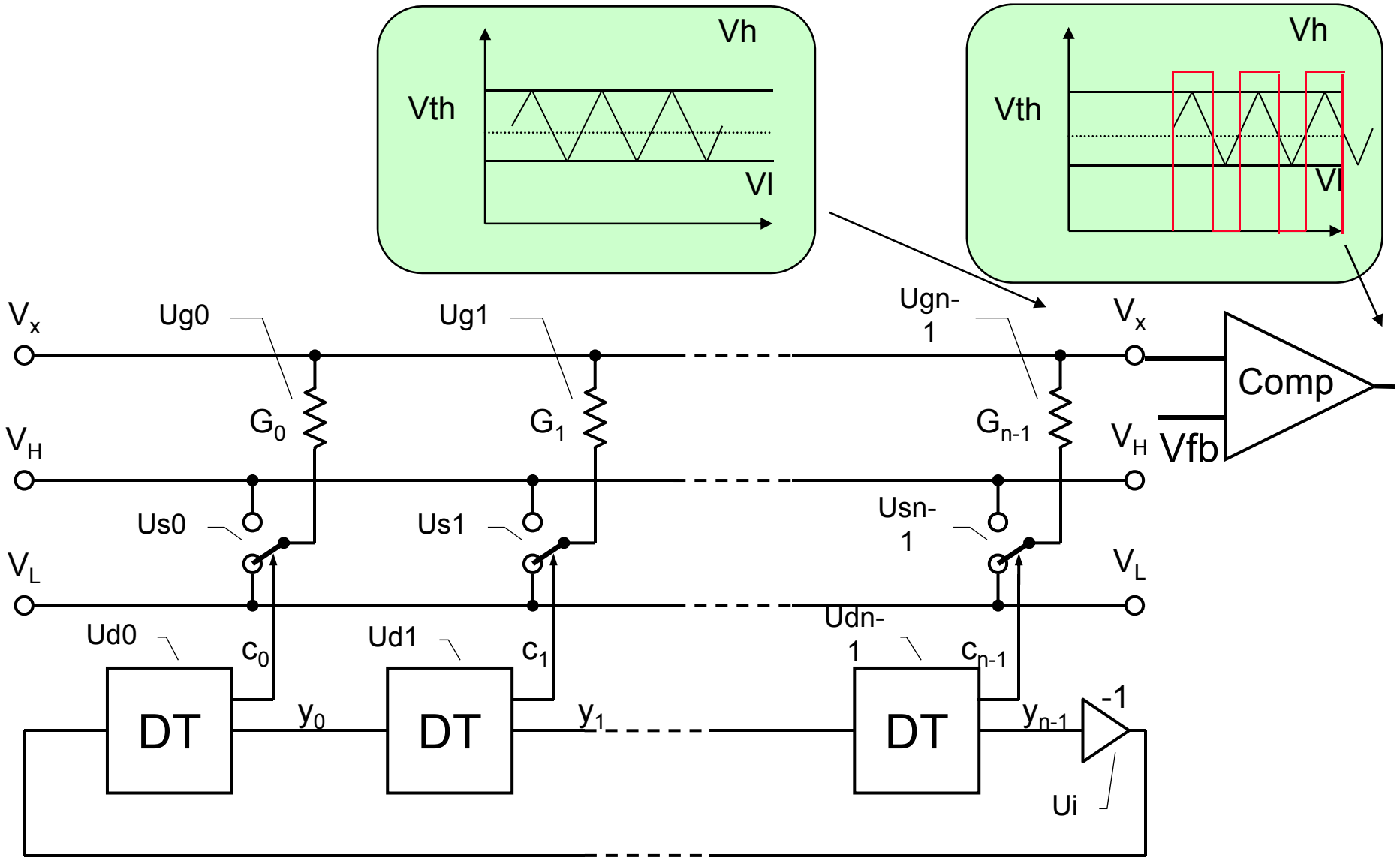


Ref: G. Schrom, F. Paillet, J. Hahn, "A 60 MHz 50W fine-grain package integrated VR powering a CPU from 3.3V", *APEC Special Presentation*, APEC 2010, Palm Springs, CA.

Bridge Driver & Current Sensor



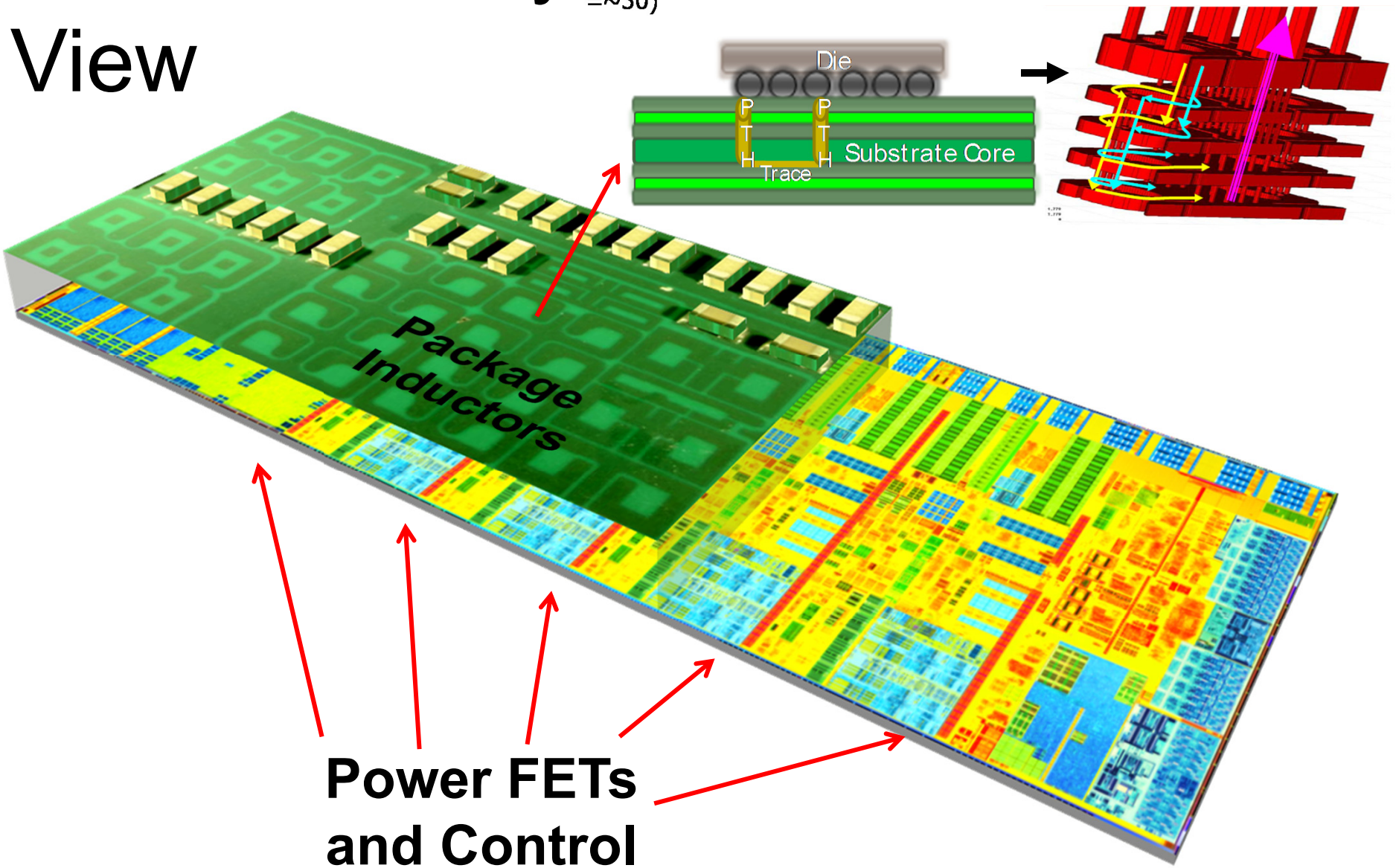
Triangular Wave Synthesizer



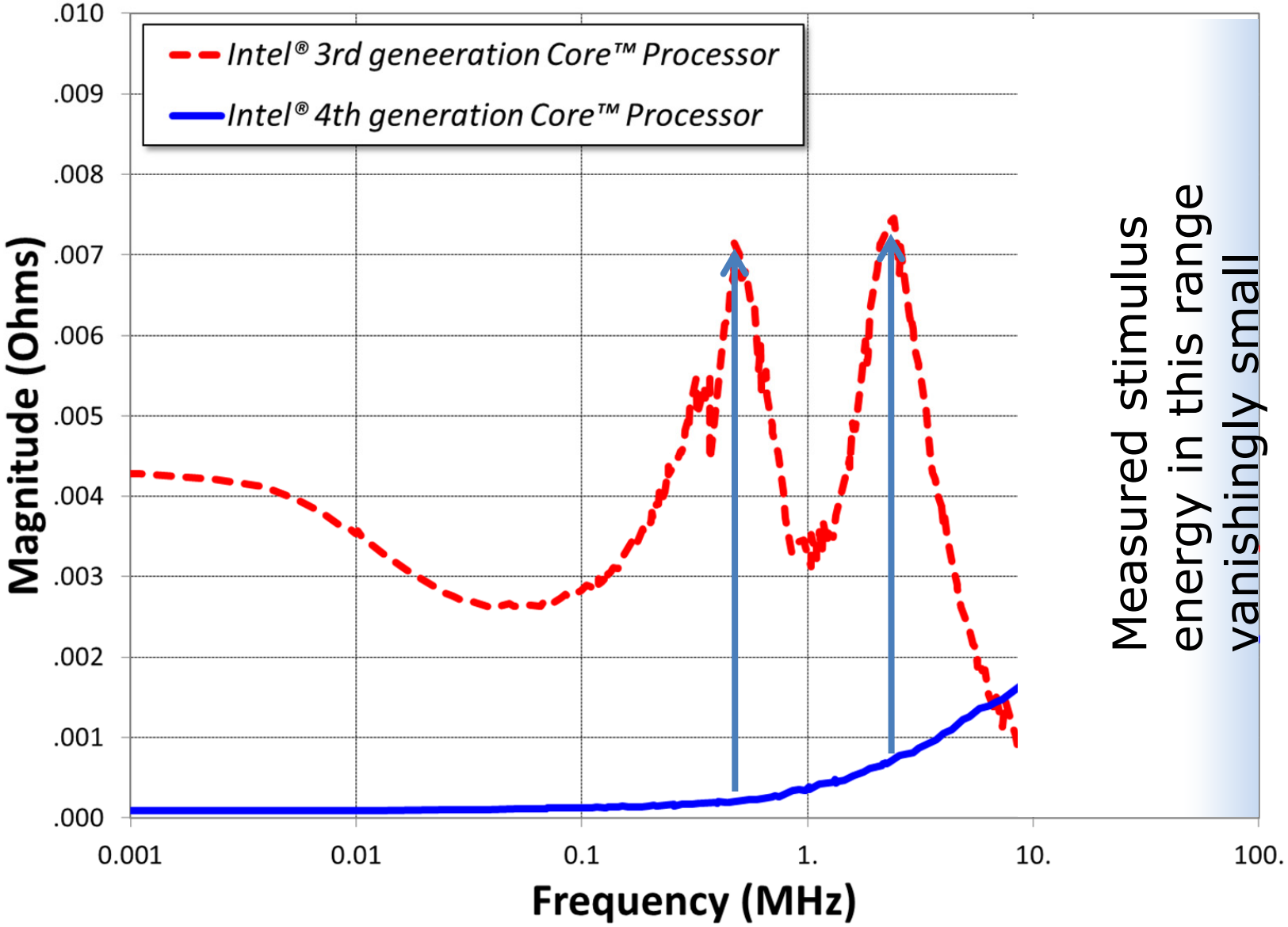
FIVR Cutaway View

Air Core Inductors (ACI): Use pkg traces and PTH's for inductor (High Q =~30)

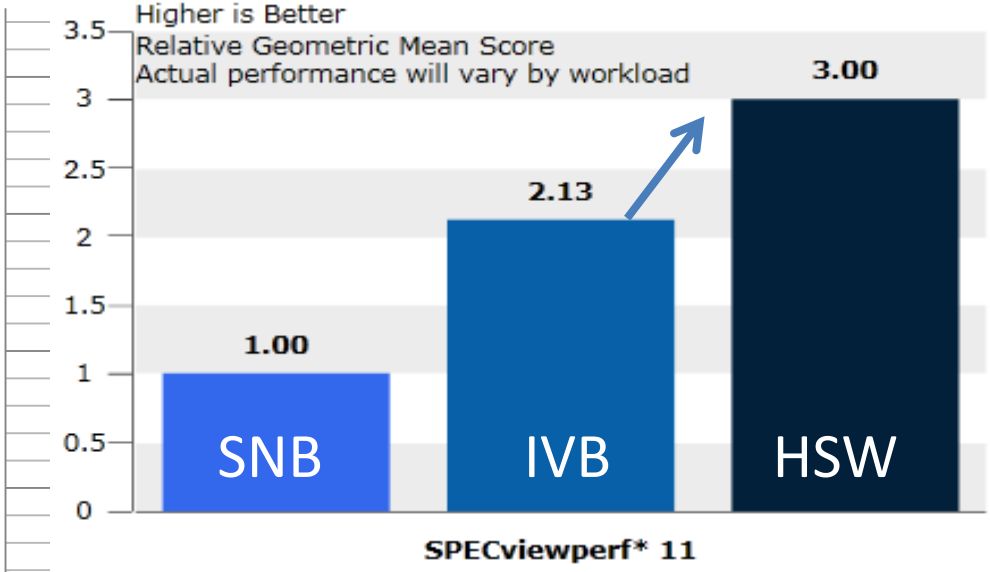
ACI



Impedance Profile

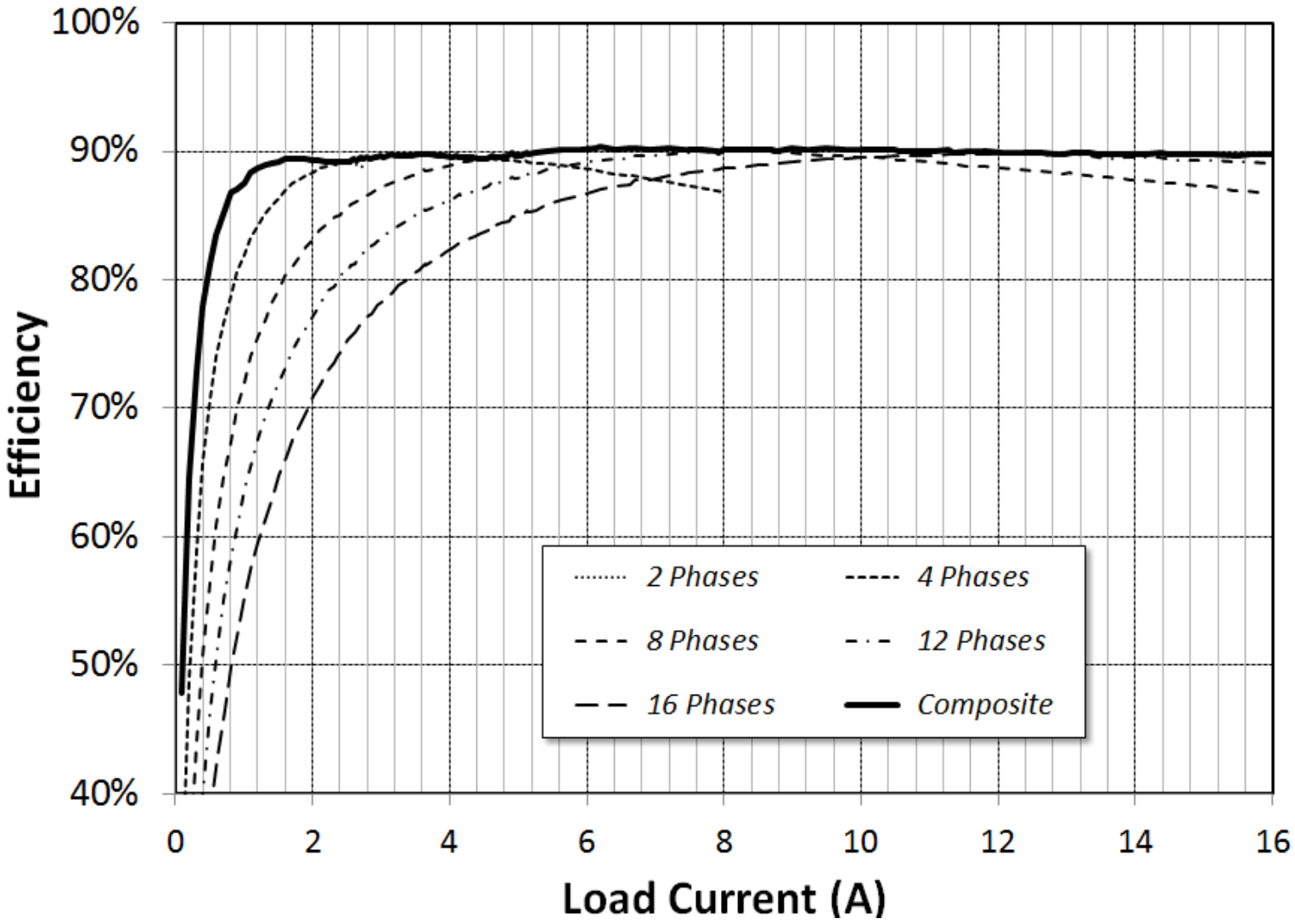


Lower Power, Higher Speed

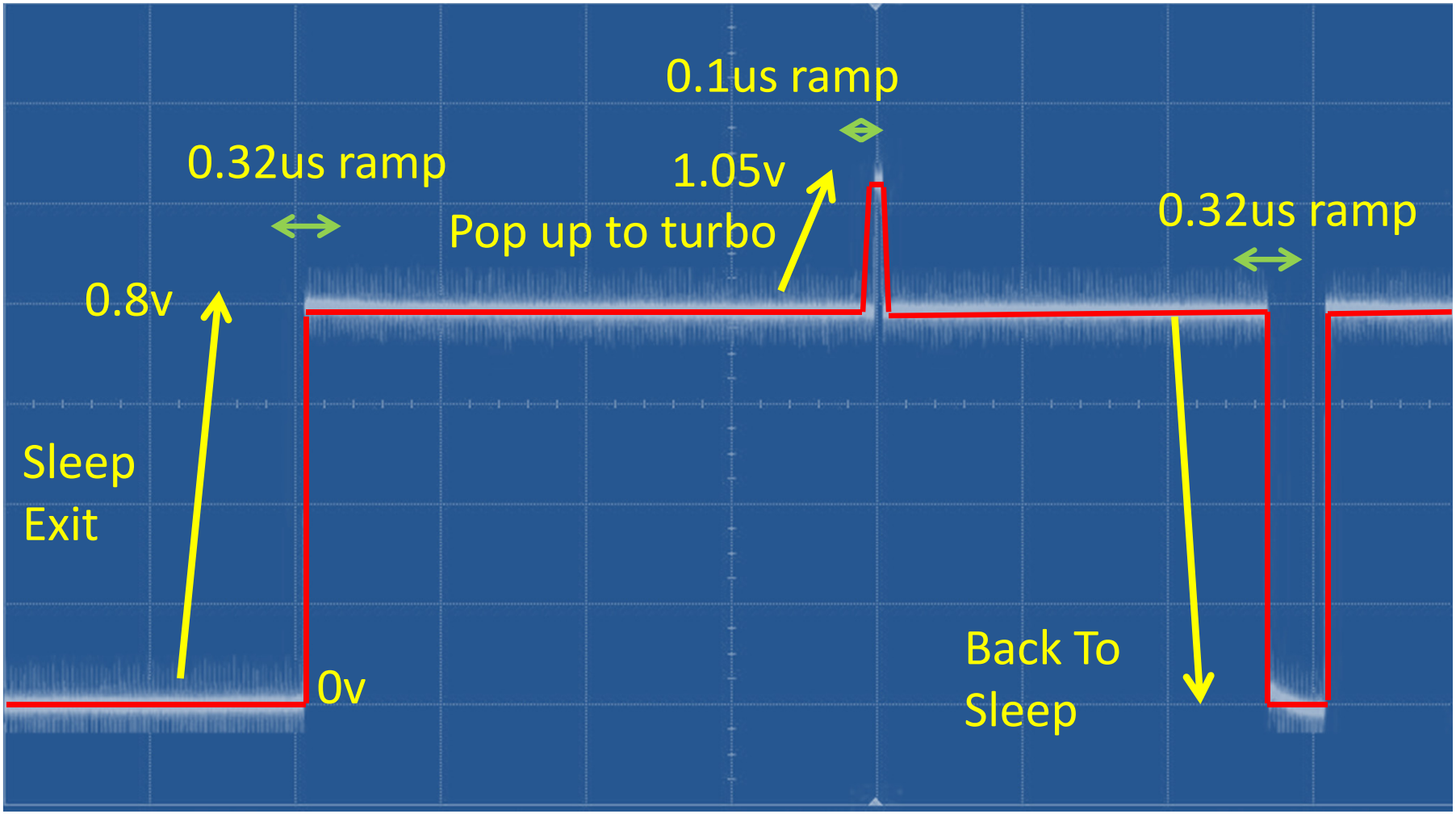


FTC Disclaimer: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information, go to www.intel.com/performance.

90% Efficiency at Full Vout



Example Fast Power Ramp

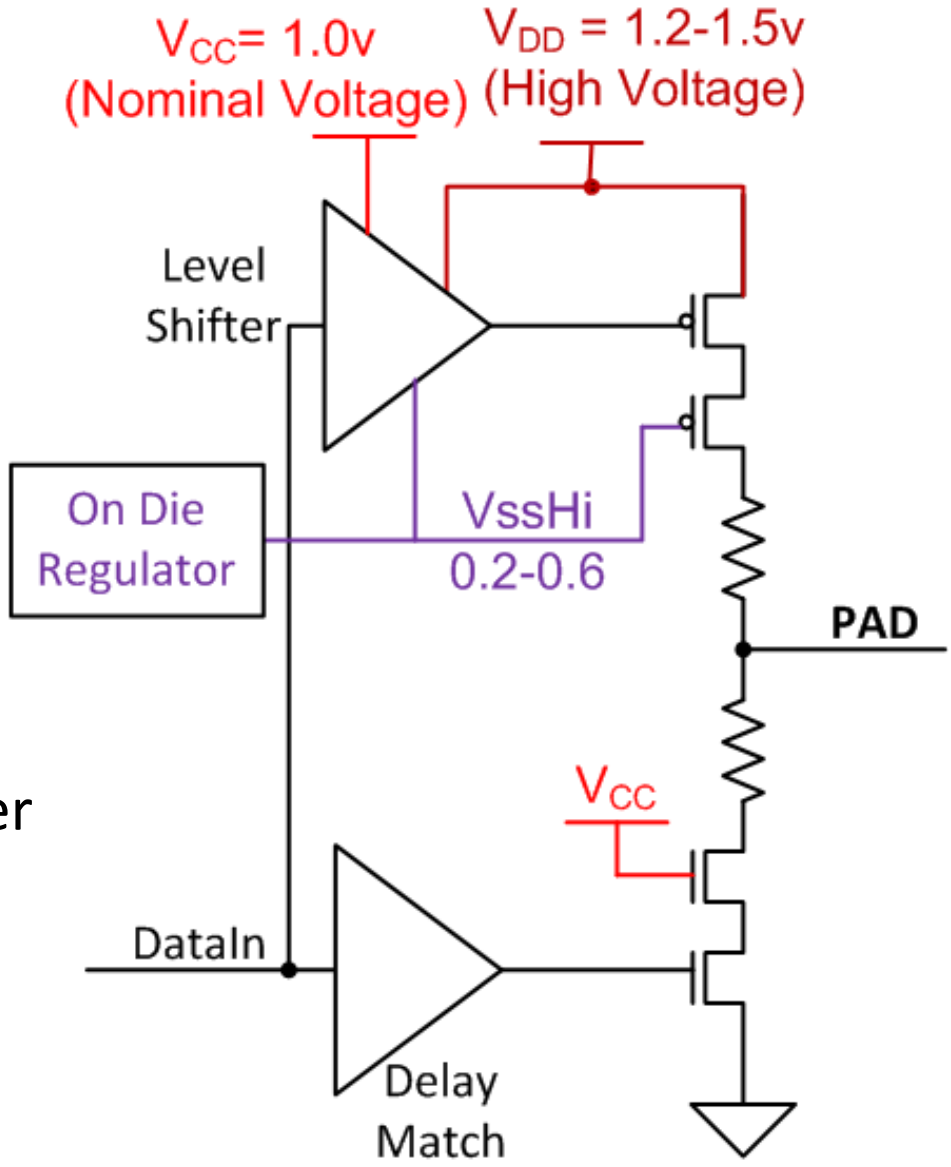


Outline

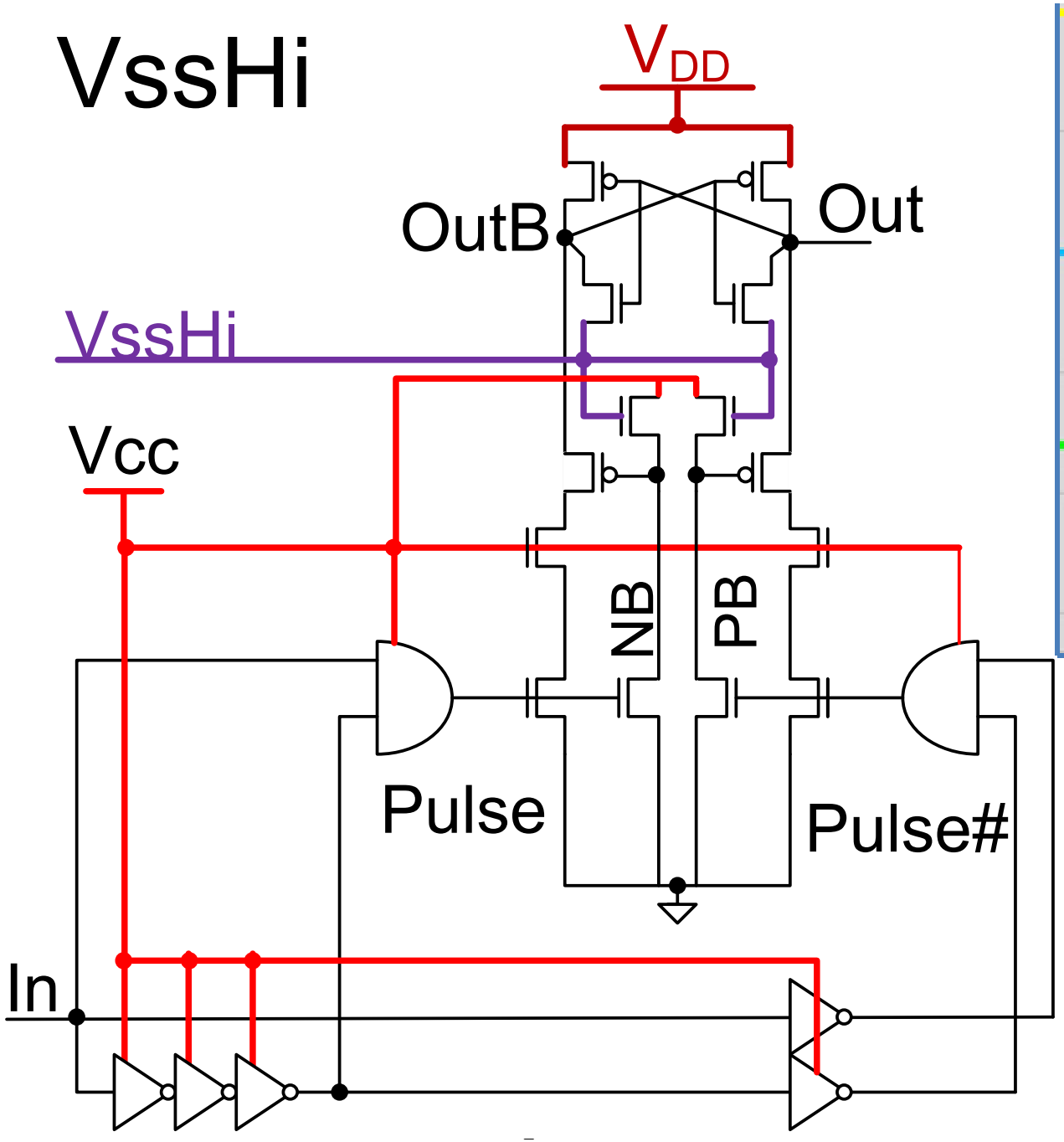
- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

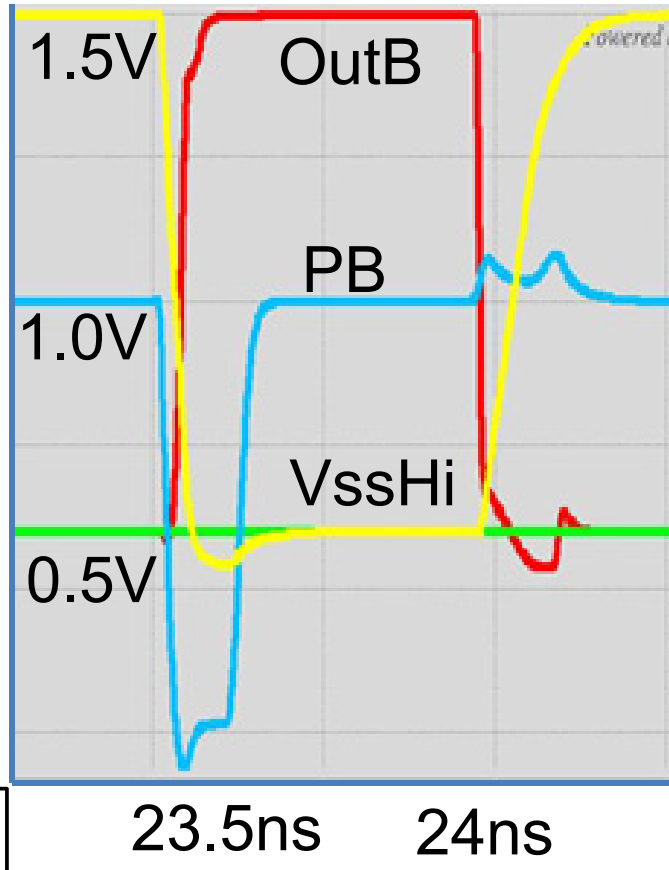
Haswell DDR

- Support 2x64 DDR3/3L/LPDDR3
  - Wide nominal voltage range from 1.2-1.5v
  - Built using native 1v
- Main focus was lower power
  - 40% Active Power, 100x Idle Power
  - Active Power: VssHi level shifter
  - Active-Idle Power: Weak Lock DLL
  - Idle Power: Stacked Power Gates



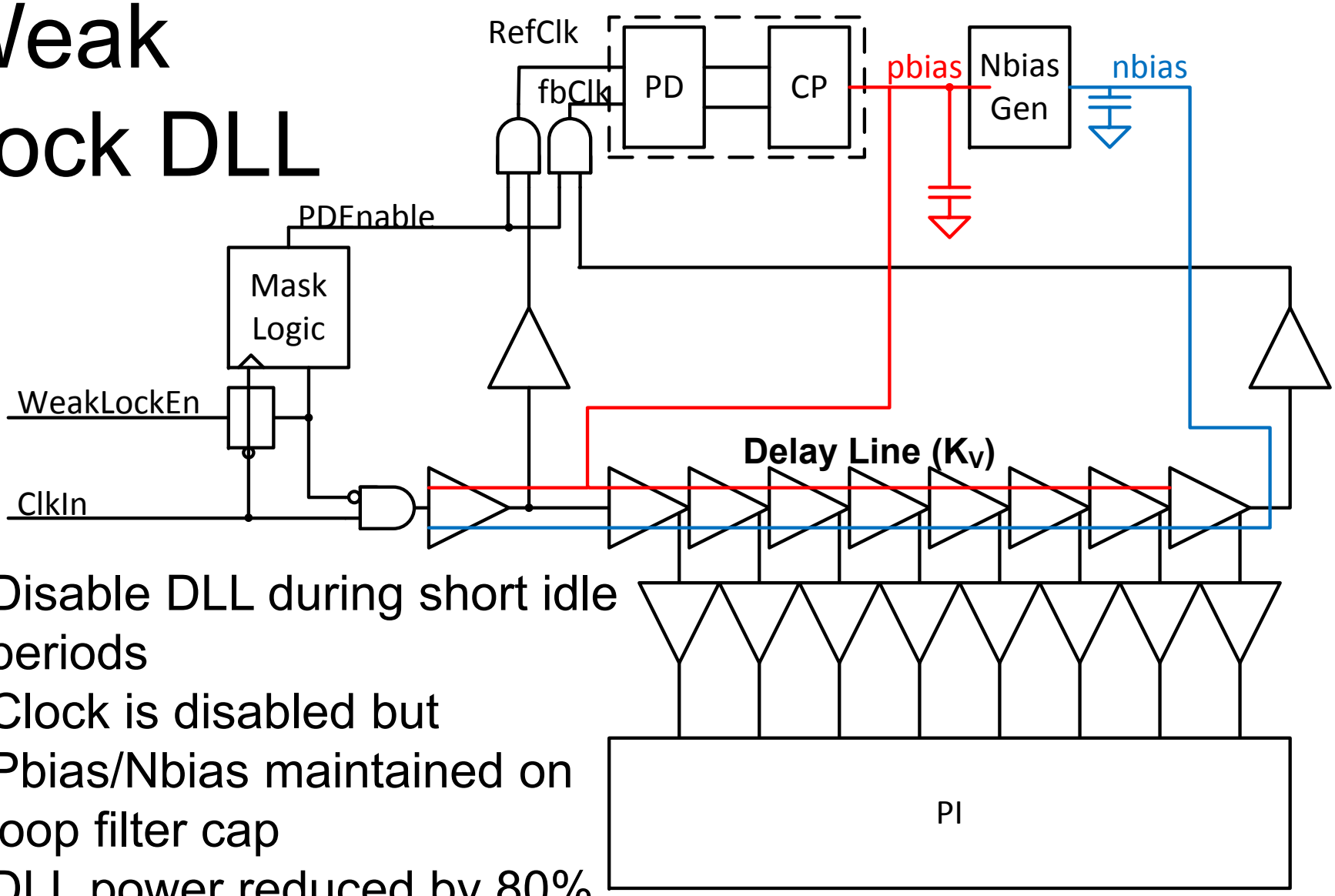
VssHi





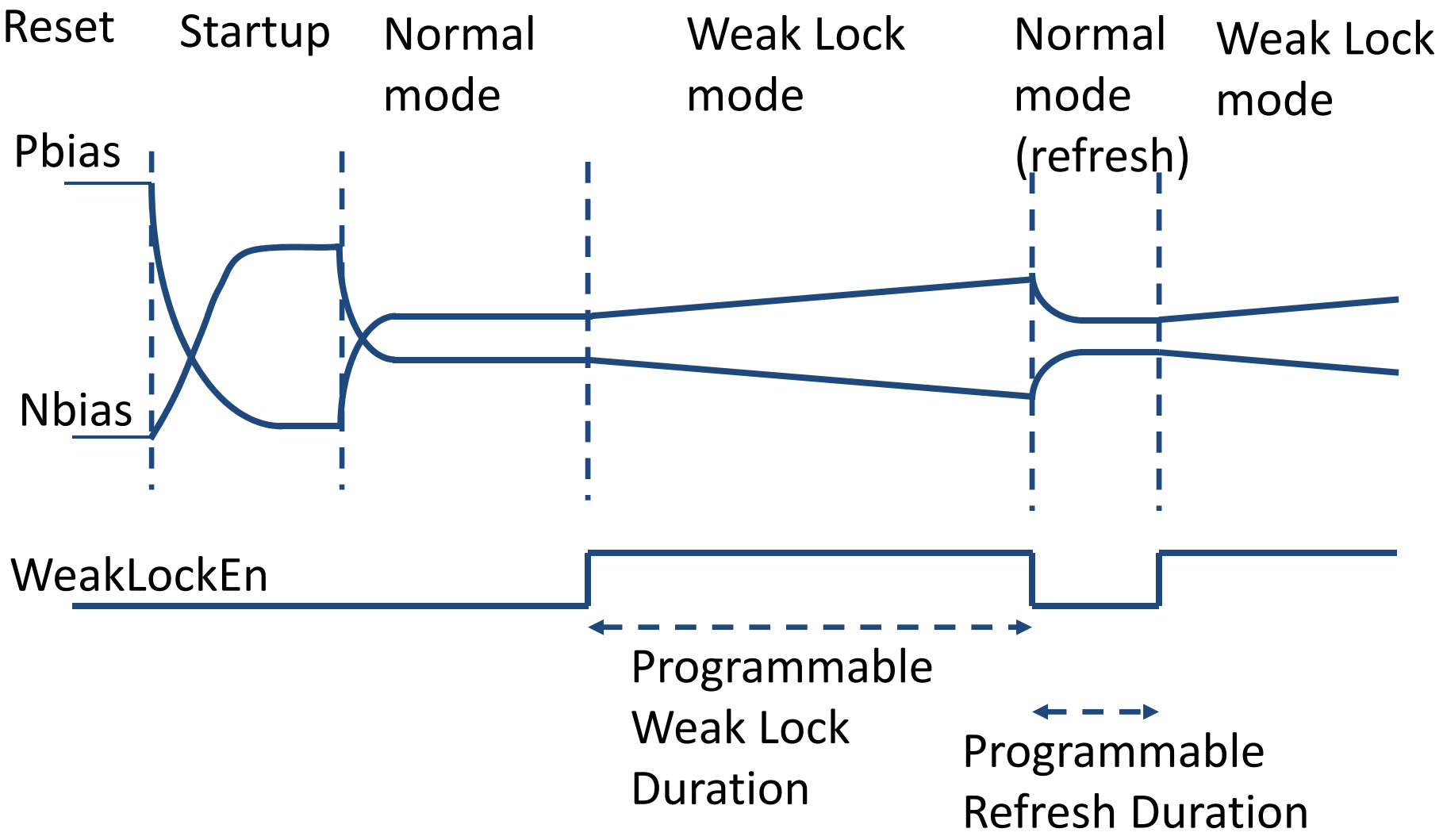
DDR VssHi
Level Shifter

Weak Lock DLL



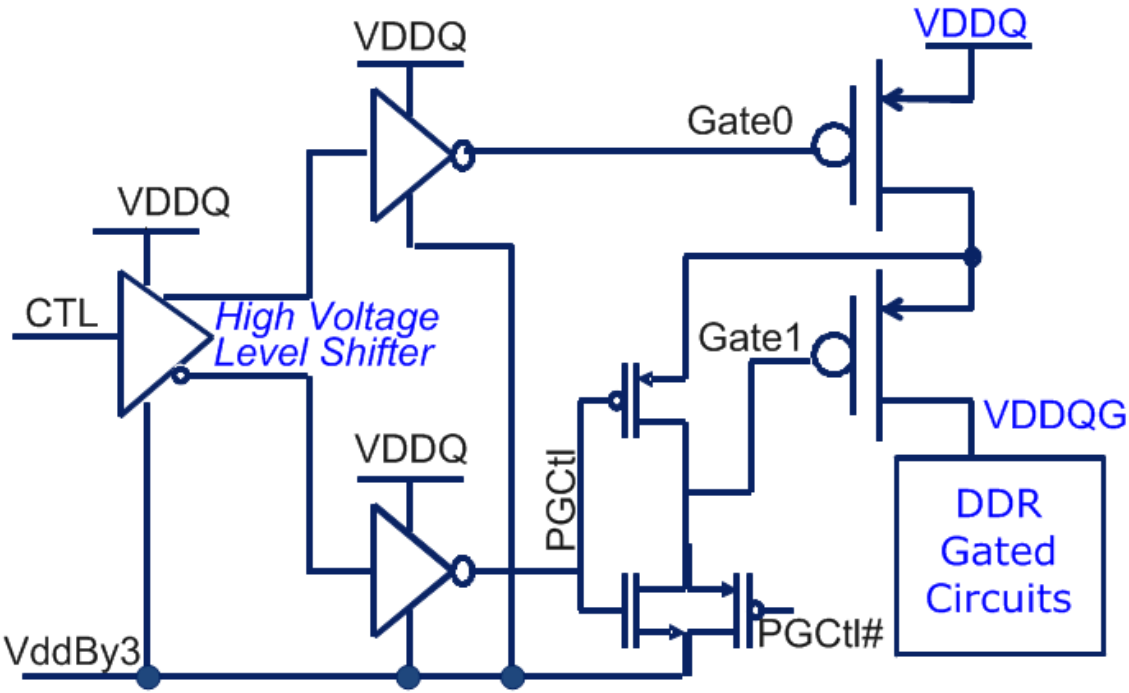
- Disable DLL during short idle periods
- Clock is disabled but Pbias/Nbias maintained on loop filter cap
- DLL power reduced by 80% with wake up time of a few nS

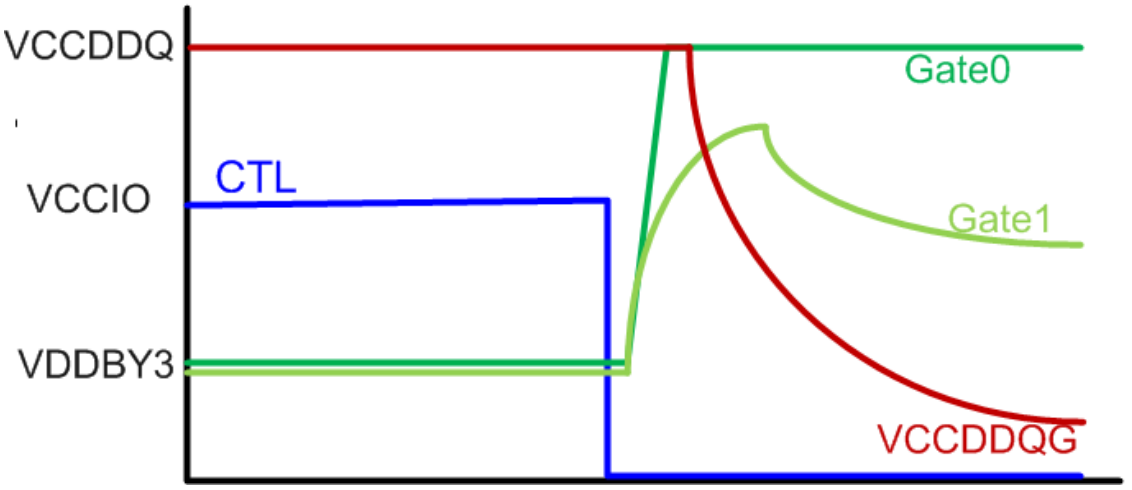
Weak Lock + Periodic Refresh



DDR Power Gate

- Stacked power gate reduces leakage by 100x vs. prior generation
- Bias voltages created by 100nA-1uA OpAmps
- ESD diodes & clamps on VddqG, next to IO buffer
- ESD Clamp RC Timer on Vddq to enable fast power ramp < 100nS





Outline

- Processor Overview
- eDRAM Integration
- On Package IO (OPIO)
- Power management
- Fully Integrated Voltage Regulator (FIVR)
- DDR
- Summary

Summary

- Haswell span full product range from fan-less to DT
- Enable 50% or more improvement in battery life
- Enable sleeker form factors and lower cost
  - Integrate up to 13 voltage regulators (FIVR)
  - Integrate PCH & eDRAM
  - Improved integrated graphics

Acknowledgement

- The authors gratefully acknowledge the work of the talented and dedicated Intel teams that have brought the Haswell family of processors to high volume Production

Please attend the Haswell demo session